VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Systém pro hromadné stahování obsahu stránek a provádění plánovaných aktivit
Web crawler system with automated action progress

2009 Tomáš Janečko

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně.

Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě dne …… ….....

Podpis autora

Poděkování

Děkuji vedoucímu bakalářské práce Ing. Radoslavu Fasugovi za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

Abstrakt

Hlavní část této bakalářské práce se zabývá návrhem systému, který bude zpracovávat internetové stránky a ukládat získaná data do výsledného souboru typu XML. Úvod pojednává o stávajícím řešení webových robotů a popisuje základní princip procházení hypertextových odkazů. Od kapitoly 3 je již pozornost věnována implementovaným funkcím a analýze vytvářeného řešení. Detailněji je zde popsána realizace hledání hypertextových odkazů, a také automatické generování nových URL. Druhá část této práce (začíná v kapitole 5) se zabývá analýzou simulováním chování reálného uživatele v rámci internetových aukcí a provádění plánovaných aktivit. Jsou zde také nastíněny problémy, které s tímto řešením souvisí.

Klíčová slova

HTML, HTTP, medaile, mince, načítání stránek, parsování dokumentu, prolézání webu, webový robot, XML, XSL transformace

Abstract

The main part of this bachelor thesis documents concept of system, which parse web pages and save retrieves data into destination file, which type is XML. Opening handle about current solution of web robots and describes passing on hyperlinks. From chapter 3 is attention devoted to implemented functions and analysis of created solution. Rich in detail is here described realization of searching hyperlinks, as well automatic generating new URL. The second part of thesis (from chapter 5) handle with analysis of simulation behaviour of real user in terms of internet auction-sales and transaction scheduled activities. Here are also foreshadoweds problems, which cohere with this solution.

Keywords

coin, fetching, HTML, HTTP, medal, parsing, web crawling, web robot, XML, XSL transformation

Seznam použitých symbolů a zkratek

ASCII – American standard code for information interchange

FIFO – First in, First out

FTP – File transfer protocol

HTML – Hypertext markup language

HTTP – Hypertext transfer protokol

HTTPS – Hypertext transfer protokol secure

IP – Internet protocol

MITM – Man in the middle

PDF – Portable document format

RSS – Rich site summary

RTF – Rich text format

SID – Session ID

SSL – Secure socets layer

TELNET – Telecommunication network

TLS – Transport layer security

TXT – Text file

UCS – Universal character set

UDP – User datagram protokol

URL – Uniform resource locator

UTF – UCS Transformation format

WWW – World wide web

XHTML – Extensible hypertext markup language

XML – Extensible markup language

XPath – XML path language

XQuery – XML query language

XSL – Extensible stylesheet language

XSLT – Extensible stylesheet language transformation

Seznam příloh

Obsah

1. Úvod

S rozšiřujícími se možnostmi a využitím internetu stále narůstá obsah dat v síti WWW, která jsou uložena v rozličných datových strukturách. Pro sběr informací se v dnešní době využívají weboví roboti, kteří mají za cíl shromáždit, co největší počet dat pro jejich pozdější zpracování. K nejvýznamnějším webovým robotům patří Google Crawler, FAST Crawler a WebCrawler.

Výsledkem této práce nemá být pouze webový robot pro sběr informací, ale systém, který bude konkretizován na základě potřeb uživatele pro získání specifických dat.

Hlavním úkolem je tedy poskytnout nástroj, který bude schopen na hypertextových dokumentech s jednotnou formou použít určenou šablonu. Důležitou vlastností tohoto zpracování by měla být schopnost aplikování tohoto řešení na internetových stránkách s různými zdrojovými daty.

Systém by měl umět pracovat na základě obecného vzoru s více stránkami se stejnou strukturou. V případě změny struktury kontrolovaných hypertextových dokumentů bude možné reagovat pouhou modifikací vstupní šablony.

Tento systém také implementuje dva různé způsoby procházení hypertextových dokumentů. První je založen na principu webového robota a postupném procházení hypertextových odkazů. Druhý způsob je realizován pomocí automatického generování URL.

V druhé části bakalářské práce bude analyzováno simulované chování reálného uživatele v rámci internetu. Do této oblasti patří automatické přihlašování a provádění plánovaných aktivit.

Hlavní důraz bude kladen na řešení zabezpečeného přihlášení protokolem HTTPS a pozdější stále udržování relace mezi uživatelem a webovým serverem. V analýze se práce bude zabývat otázkou, jak řešit na internetových aukcích průběh automatického příhozu s ohledem na vykonání požadavku na straně klienta.

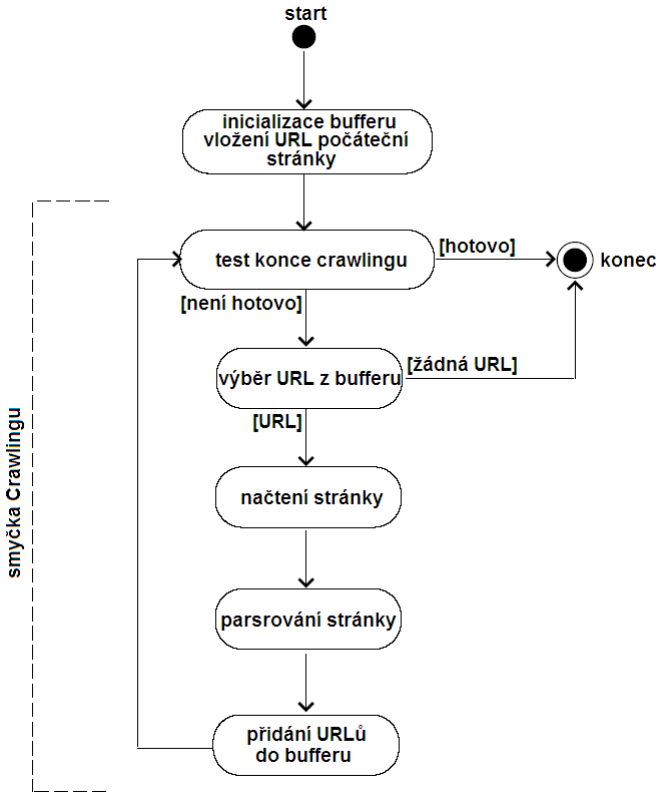
2. Stávající řešení v oblasti stahování internetového obsahu

Pro stahování internetového obsahu se používají automatizované prostředky, které pracují na základě předem stanovených podmínek. Při velkém rozsahu dnešních počítačových sítí je kladen největší důraz na rychlost a korektnost potřebných výsledků. V dalších kapitolách je detailněji rozebrán současný způsob řešení a nasazení vytvořeného systému.

2.1 Webovi roboti

Webový robot je program, který prochází celosvětovou síť WWW. Jeho hlavní úkolem je vytvoření kopie navštívených stránek pro jejich pozdější zpracování. Roboti [1] mohou nacházet využití k provádění automatických údržbářských prací na internetových stránkách, jako jsou kontrola funkčnosti odkazů nebo validace HTML kódu. Tyto programy mohou být napsány také k získávání specifického druhu informací z webových stránek, kdy se například snaží zjistit e-mailové adresy.

Webový robot musí mít nejen dobře zvolený algoritmus (obrázek 1), ale také velice optimalizovanou architekturu. K prováděným akcím patří ukládání URL v kanonizovaném stavu nebo zpracovávání načtené stránky a extrakce informací potřebných pro robota tzv. parsing.



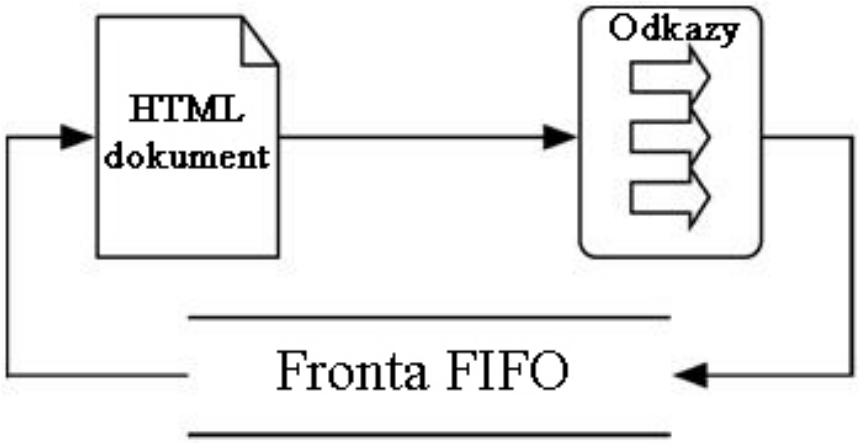
Obrázek 1: Obecný algoritmus webového robota

2.2 Algoritmy webových robotů

Zvolený algoritmus webového robota je jedna z nejdůležitějších vlastností pro zpracování internetového obsahu. Přístup a zvolený způsob řešení má zásadní vliv na obdržená data, jejich kvalitu, rychlost práce webového robota a efektivitu procházení dokumentů.

2.2.1 Breadth-first

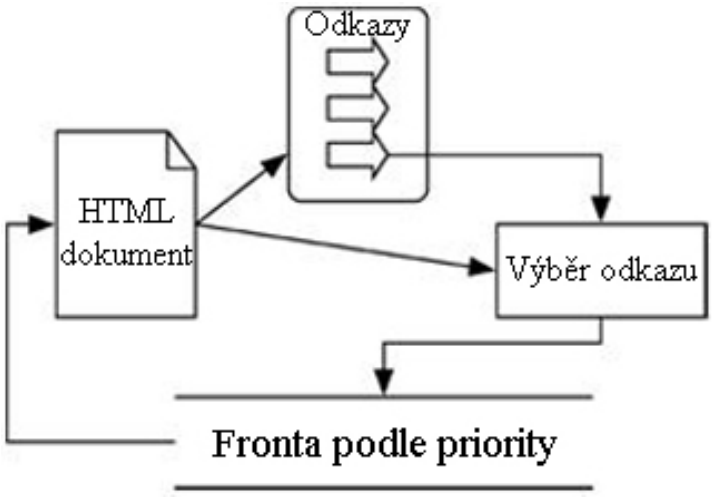
Tento algoritmus obsahuje nejjednodušší strategii k procházení internetových dokumentů. Způsob algoritmu byl prozkoumán již v roce 1994 a jeho základní myšlenka spočívá v tom, že webový robot obsahuje FIFO frontu, ze které kontroluje odkazy v přesném pořadí, v jakém jsou uloženy. Pokud je fronta plná, tak webový robot ukládá pouze jeden nový platný odkaz z aktuálně zpracovávané internetové stránky. Breadth-first algoritmus (obrázek 2) je znázorněn níže. Řešení pomocí tohoto algoritmu je demonstrováno jako základ webových robotů.



Obrázek 2: Webový robot Breadth-first

2.2.2 Best-first

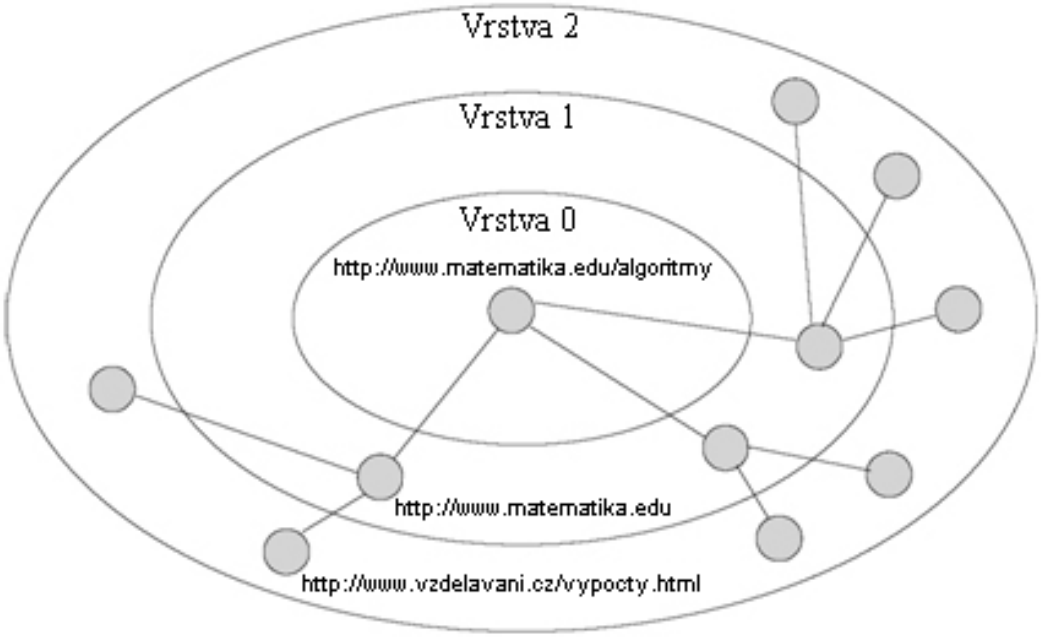
Základní myšlenkou je, že webový robot obsahuje frontu hypertextových odkazů a pro kontrolu dalšího dokument se vybírá odkaz, který nejlépe splňuje podmínky podle upřesněného odhadu. Rozdílná řešení Best-first [2] algoritmu implementující odlišnou složitost a efektivitu mohly být napsány pouze na základě sofistikovanějších a komplikovanějších kritériích pro výběr odkazů. Algoritmus v provedení Naive Best First (obrázek 3) využívá proces výběru relevantního odkazu za pomoci počítání slovníkové podoby mezi tématy, klíčovými slovy a obsahem stránky. URL s nejlepším odhadem a podobností je pak vybrána jako další v pořadí pro procházení a kontrolu.



Obrázek 3: Webový robot Best-first

2.2.3 Context focused

Toto řešení pracuje na způsobu klasifikace stránek podle kategorií nebo konkrétních tematických okruhů jako algoritmus Focused [3]. Na začátku webový robot požaduje zadání tematické oblasti a snaží se naleznout stránky odpovídající těmto požadavkům. Myšlenka je tedy následující. Hledáme-li informace například o numerické analýze, tak půjdeme nejprve na matematickou domovskou stránku. Context focused algoritmus ke klasifikaci navíc využívá výpočet vzdálenosti mezi URL zpracovávané stránky a relevantními URL. V tomto případě se využívá tzv. kontextového grafu (obrázek 4) rozděleného do vrstev.



Obrázek 4: Kontextový graf

2.3 Dostupná řešení na trhu

Stávající možnosti stahování online obsahu jsou velice rozmanité a nabídka programů, které jsou schopny tuto funkci zastat také. Koncový spotřebitel má na výběr, zda si zvolí variantu nabízenou zdarma nebo jestli využije placené produkty. Vybraný program by měl splňovat konkrétní specifické požadavky uživatele na vlastnosti a funkcionalitu. Zde jsou popsány principy, které by měly být dodrženy.

Seznam souborů určených pro stahování

Nabízený produkt by měl nabídnout možnost definovat seznam všech doménových jmen, ze kterých bude probíhat stahování. Stanovení povolených nebo zakázaných souborů či adresářů. Určit seznam souborů a jejich koncovek, který si uživatel přeje nebo nepřeje stahovat. Zadáním hloubky zanoření stahování zajistit relativní vzdálenost vzhledem k první adrese URL. Z dalších nastavení to může být stanovení limitních hodnot souborů, kdy je uživatel schopen definovat minimální nebo maximální velikost stahovaných souborů.

Sledování odkazů

Stahování souborů může být omezeno pouze na seznam domén, popř. stahování může být omezeno pouze na právě stahovanou doménu. Každý webový server také zpřístupňuje informace na konkrétním portu. Na tomto portu je pak většinou poskytován celý obsah. Přechod na jiné porty je opět možné zakázat, popř. povolit například konkrétní seznam portů. Webový robot podporuje určitou sadu protokolů, na kterých je schopen se službami na internetu komunikovat. Takový standardní robot podporuje protokol HTTP. Další možné protokoly, které některé ze zmiňovaných nástrojů podporují, jsou např. HTTPS či FTP.

Způsob stahování

Webový robot může stahovat stránky sériově, tedy postupně za sebou. Toto řešení však není optimální, protože např. při stahování již dvěma vlákny může dojít k navýšení rychlosti stahování jednoho souboru i o desítky procent. Stahování je tedy možné paralelizovat a to buď tak, že se jeden soubor bude stahovat několika vlákny, popř. několik různých souborů bude stahováno souběžně. Další možností paralelního stahování je získávání souborů z několika zrcadel současně, tzn. jeden stejný soubor je uložen na více serverech. Možnosti datové komprese jsou, kdy program může s požadavkem zaslat na server také informaci o této své schopnosti. V případě, že server tuto kompresi podporuje, může zpět vrátit odpověď v komprimované podobě a musí o

tom informovat v zaslaných hlavičkách. Množství přenášených dat je možné tímto způsobem u textových informací výrazně snížit. Při sdílení připojení k internetu je vhodné např. omezit maximální rychlost stahování, aby linka byla volná pro ostatní uživatele.

Ukládání stažených souborů

Na výběr by měla být možnost ukládání s absolutními nebo relativními cestami odkazů. Při ukládání HTML souborů je možnost uložení souborů s relativními cestami vhodnější pro budoucí přesun celého adresáře se staženým webem. Během stahování souborů je často nutné soubor před uložením upravit tak, aby lokální kopie odkazovala na současně stažené soubory. Občas je ale vhodné zachovat původní verzi stahovaného souboru.

Procházení HTML stránek a vyhledávání odkazů

Pro koncového spotřebitele je velkou výhodou, když má k dispozici možnost definování vlastních elementů. Na kontrolované HTML stránce existují případy, kdy je vhodné zadat určité entity, které budou při procházení dokumentů zpracovávány. Na některých webových stránkách jsou umístěny formuláře pro vyhledávání nebo přihlášení. V případě zadání konkrétního výrazu do vyhledávače, je poté možné stáhnout vyhledané stránky na základě vygenerovaného odkazu včetně skrytých polí ve webovém formuláři. Každý webový robot disponuje sadou algoritmů pro zpracování různých typů souborů. Standardem jsou HTML stránky. Mezi další typy, které závisí na možnostech robota, jsou: zpracování kaskádních stylů (.css), javascriptů (.js), získávání odkazů z flashe (.swf), popř. javovských (.class) souborů a mnohé další.

Připojení na server

V řadě podnikových sítí se k přístupu na internet používá proxy server. Může se jednat o transparentní proxy, na kterou se není nutné přihlašovat, stále se ale používají i proxy servery, do kterých je nutné se přihlašovat, potom je tento parametr jediný způsob, jak požadované stránky stáhnout. Na internetu se nalézají také stránky, které jsou chráněny heslem a bývají zabezpečeny pomocí přístupové autentizace. V případě podpory tohoto parametru ze strany robota je přístup na tyto stránky možné uskutečnit po zadání uživatelského jména a hesla.

Rozšířené funkce

Většina webových robotů nabízí stahování obsahu na základě daného seznamu pravidel. V případě, že potřebujeme zpracovávat stránky, které jsou specificky číslované nebo jinak systematicky označené, je možné, v případě podpory tohoto parametru, takový seznam stránek automaticky vygenerovat. K dalším speciálním možnostem můžeme zařadit zobrazení statistik stahování souborů z webového serveru.

Porovnání dostupných programů

Výsledky s funkcionalitou a možnostmi stávajících řešení jsou zobrazeny v Tabulce 1. Tabulka zobrazuje řešení stávajících dostupných programů.

Název produktu	Stahování ve vláknech	Filtr souborů	Podpora Java appletů	Proxy server	Generování URL
HTTrack Website Copier	Ano	Ano	Ne	Ne	Ano
Teleport Pro	Ano	Ano	Ano	Ne	Ne
Web Downloader	Ano	Ne	Ne	Ano	Ne
Website Extractor	Ne	Ano	Ne	Ne	Ne
Web Boomerang	Ne	Ano	Ne	Ano	Ne
Offline Commander	Ano	Ano	Ne	Ne	Ne
Get Site	Ano	Ano	Ne	Ne	Ne

Tabulka 1: Funkce programů

2.4 Důvody použití

Vytvořený systém bude využíván ke zpracování nabídky internetových obchodů, stahování online aukcí a uložení jejich obsahu do požadované struktury. Veškerá získaná data jsou uložena v souborech typu XML. Výhodou tohoto řešení je, že všechny uložené informace mohou mít jednotné pojmenování a tím se usnadní práce a přístup k výsledným souborům. S takto vytvořenými strukturami lze dále pracovat.

Řešení by mělo být odlišné od stávajících dostupných programů, které jsou schopny stahovat pouze obsah z jediného typu serveru. Hlavní příčinou tohoto problému je jejich pevně daná struktura, která musí být dodržena. V opačném případě se stává program nepoužitelným.

Velká část dat v rámci e-shopů a internetových aukcí, která jsou prezentována na webových stránkách, je automaticky generována systémem, který vytváří výsledný vzhled internetových stránek a jejich celkovou strukturu.

Vyvíjené řešení není omezeno na jeden typ datových struktur, ale lze ho nadefinovat obecně pro libovolnou strukturu. Z toho vyplývá, že pokud se změní například systém, který internetové stránky vytváří a jejich výsledná generovaná forma bude jinak definována, tak vytvářený systém umí na tuto změnu reagovat modifikací zadáním, jak má s webovou stránkou pracovat.

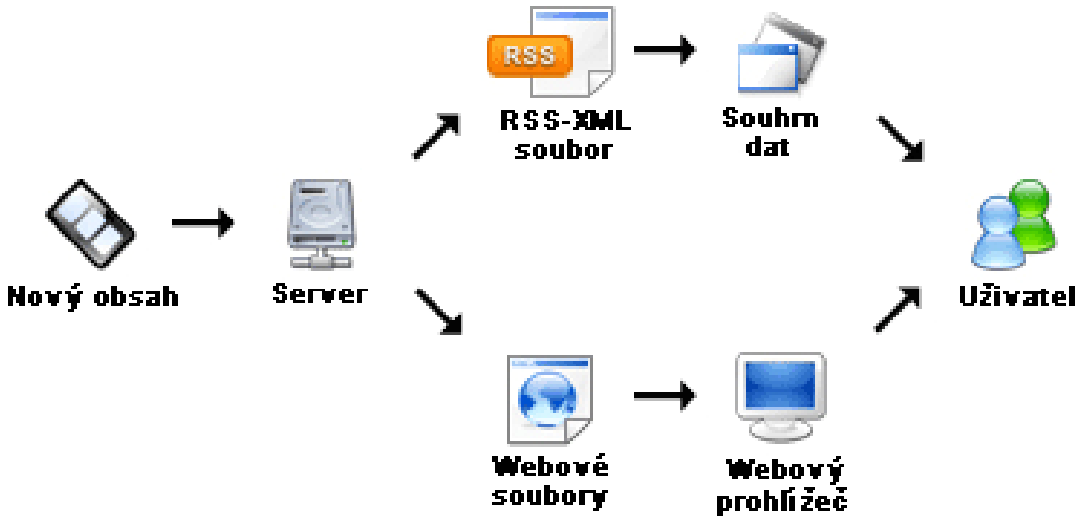
2.4.1 E-shopy

Velice využívanou metodou internetových obchodů k uvedení svých produktů a informování spotřebitelů o nabízených novinkách je zavedení a používání RSS. Jedná se o informační zdroje, které jsou schopny automaticky uživatele upozorňovat na nové zprávy, přidané informace a další novinky.

Technologie RSS (obrázek 5) umožňuje uživatelům Internetu přihlásit se k odběru novinek z webu, který nabízí RSS zdroj. Tento zdroj se většinou vyskytuje na stránkách, kde se obsah mění a přidává velmi často.

RSS řeší problém lidí, kteří pravidelně využívají webové stránky. Uživateli to dovoluje zůstat informovaný a získávat poslední obsah ze stránek, o které se zajímá. Koncový spotřebitel ušetří velké množství času, který by musel věnovat stálé kontrole aktuálnosti dat.

Vytvářený program dokáže tyto RSS zdroje [4], které jsou uloženy ve formátu XML přečíst, přetransformovat a uložit do podoby, kterou uživatel zadal. Všechna výsledná získaná data jsou uložena v jednotné struktuře datového typu XML dokumentu.



Obrázek 5: Způsob využití RSS

S takto vytvořenými strukturami lze dále pracovat. Jsme schopni zjistit celkovou nabídku a sortiment produktů konkrétního internetového obchodu. Snadno také vyhledáme cenu výrobku, popis a jiné požadované informace, které jsou potřebné.

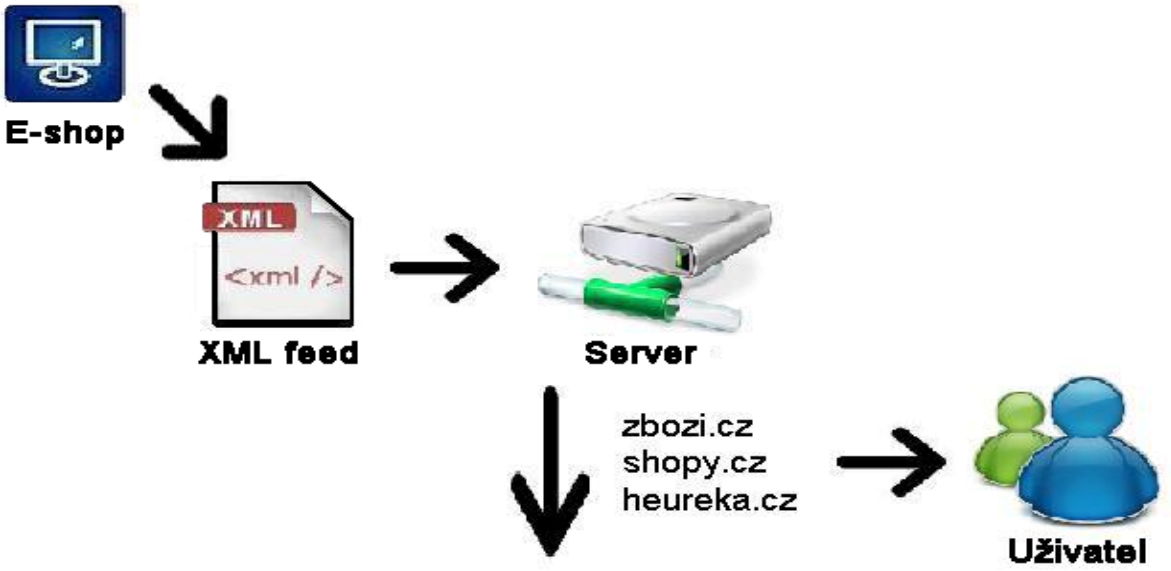
V další fázi používání systému se nabízí možnost kontrolování kvality, rozmanitosti nabídky, cenové relace a porovnání schopností internetových obchodů nabídnout zákazníkům potřebné služby.

RSS kanál (obrázek 6) nabízí a zobrazuje aktuální data na serveru, ke kterým je velice snadný přístup. Pokud uživatel využívá RSS čtečku, tak pouhou modifikací dat v tomto kanálu se okamžitě dozví o změně a stávající nabídce.

```
−<rss version="0.91">
  −<channel>
      <title>Digg Top Story RSS Feed</title>
      <link>http://www.digg.com</link>
    −<description>
        Creates a simple RSS feed that contains the top story from Digg.
      </description>
      <language>en</language>
    −<image>
        <title>Digg Top Story RSS Feed</title>
        <link>http://www.digg.com</link>
       −<url>
          http://www.kapowtech.com/images/open/comingsoon_logo.gif
        </url>
      </image>
    −<item>
        <title>SUVs Escape Guzzler Tax </title>
       −<link>
          http://blog.wired.com/cars/2006/10/suvs_escape_guz.html
        </link>
        <description/>
      </item>
    </channel>
  </rss>
```

Obrázek 6: RSS kanál

Další z možností, kterou internetové obchody využívají je generování XML feedů (obrázek 7). Takto generované dokumenty jsou tvořeny z katalogu a nabídky internetových obchodů. Z vytvořených dokumentů lze opět vyčíst veškeré informace o nabízených produktech a s využitím vytvořeného systému přetransformovat do jednotné podoby.



Obrázek 7: XML feed

2.4.2 Internetové aukce

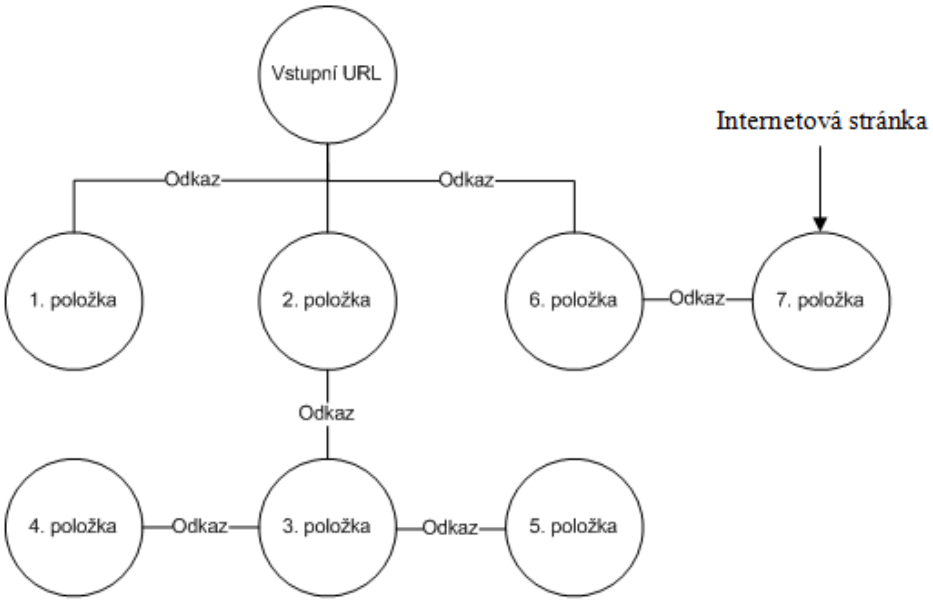
Mezi jiná využití patří procházení internetových aukcí, kdy je opět možnost uložení poskytovaného obsahu do databáze. Pro kontrolu duplicit záznamů v databázi je využito ID ve formě čísla. Případně dvojice čísel, která reprezentuje klíč záznamu. K jednoznačnému určení položky v databázi se využívá také textový řetězec, ale vzhledem k efektivitě a potřebám systému bylo vybráno řešení reprezentováno v číselné podobě.

Procházení jednotlivých webových stránek aukcí není realizováno pouze hledáním odkazů na platné dokumenty, ale také generováním URL. Získané odkazy jsou tvořeny dosazováním číselných hodnot za uživatelem označený textový řetězec v názvu URL.

Publikované výsledky aukcí se liší svou strukturou, návazností na další položky v aukci, a také na odlišné způsoby procházení celého aukčního katalogu. K detailnějšímu zpracování byly aukce rozděleny do následujících kategorií.

1a) Stromová struktura s jednou položkou

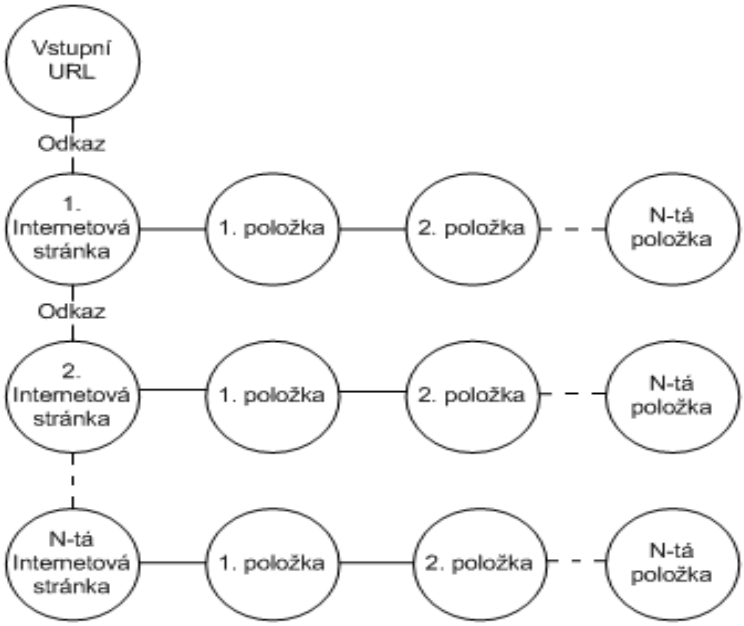
Tento typ aukce je založen na principu, kdy se konkrétní položka nachází na samostatné internetové stránce (obrázek 8) a s dalšími záznamy v aukci je propojena pomocí hypertextových odkazů. V tomto případě je nejefektivnější použít procházení pomocí webového robota se zadanými kritérii na typ odkazu, který se má následovat a dále zpracovávat.



Obrázek 8: Aukce s jednou položkou na stránce

1b) Stromová struktura s více položkami

Aukce tohoto typu je také realizována průchodem, který sleduje a vyhledává potřebné hypertextové odkazy. V tomto případě se na internetové stránce nachází seznam položek (obrázek 9), který může být reprezentován např. tabulkou. Všechny záznamy na stránce se musí jednoznačně identifikovat. Systém zajistí korektní rozdělení na základě vstupních požadavků uživatele a ve výsledném souboru uloží data jako samostatné položky.



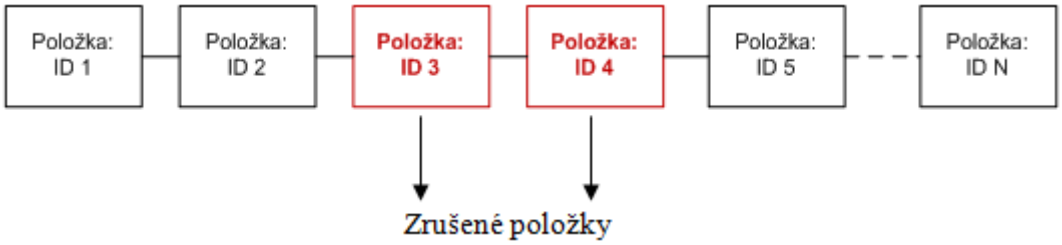
Obrázek 9: Aukce s více položkami na stránce

2) Sekvence aukcí s ID

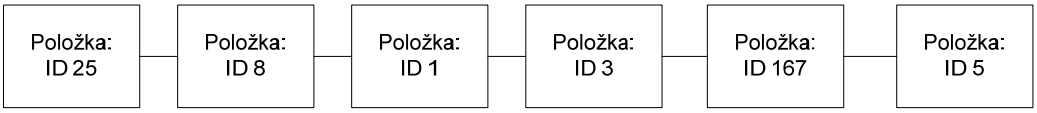
V případě aukcí, které mají položky označené jednoznačným identifikátorem, je vhodné použít metodu generování URL. Změnou tohoto identifikátoru v URL nám webový server vrátí příslušnou internetovou stránku s obsahem patřícím položce s daným ID. Sekvence položek lze generovat pomocí jednoho nebo dvou číselných identifikátorů.

2a) Sekvence ID s jedním číslováním

K hlavním problémům souvisejícím s číselnými sekvencemi je vynechání položky z důvodu zrušení aukce (obrázek10) nebo posloupnost neuspořádaného seznamu (obrázek 11).

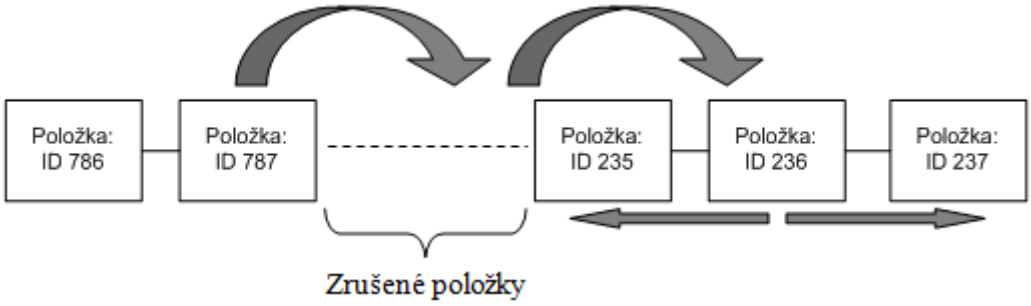


Obrázek 10: Sekvence s vynechaným ID



Obrázek 11: Neuspořádaná sekvence

Řešením těchto problémů je implementace skoků v rámci zadaného rozsahu pro generování URL. Tyto posuny (obrázek 12) v ID hodnotách položek se snaží minimalizovat dobu, po kterou by byly generovány URL a webový server by stále vracel internetové stránky neodpovídající požadavkům uživatele.

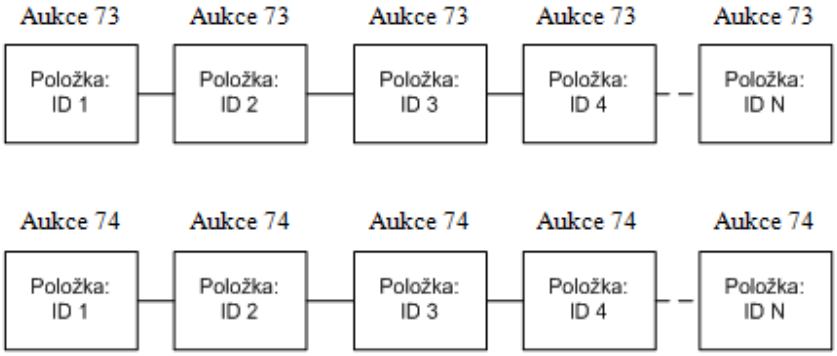


Obrázek 12: Realizace skoků v ID

Z důvodu, že nelze ovlivnit a zjistit předem přesnou položku, na kterou se skokem dostaneme, dochází k postupnému se vracení v hodnotách ID až do počátku aukce, na kterou jsme se dostali nebo k vrácení se do posledního známého místa skoku.

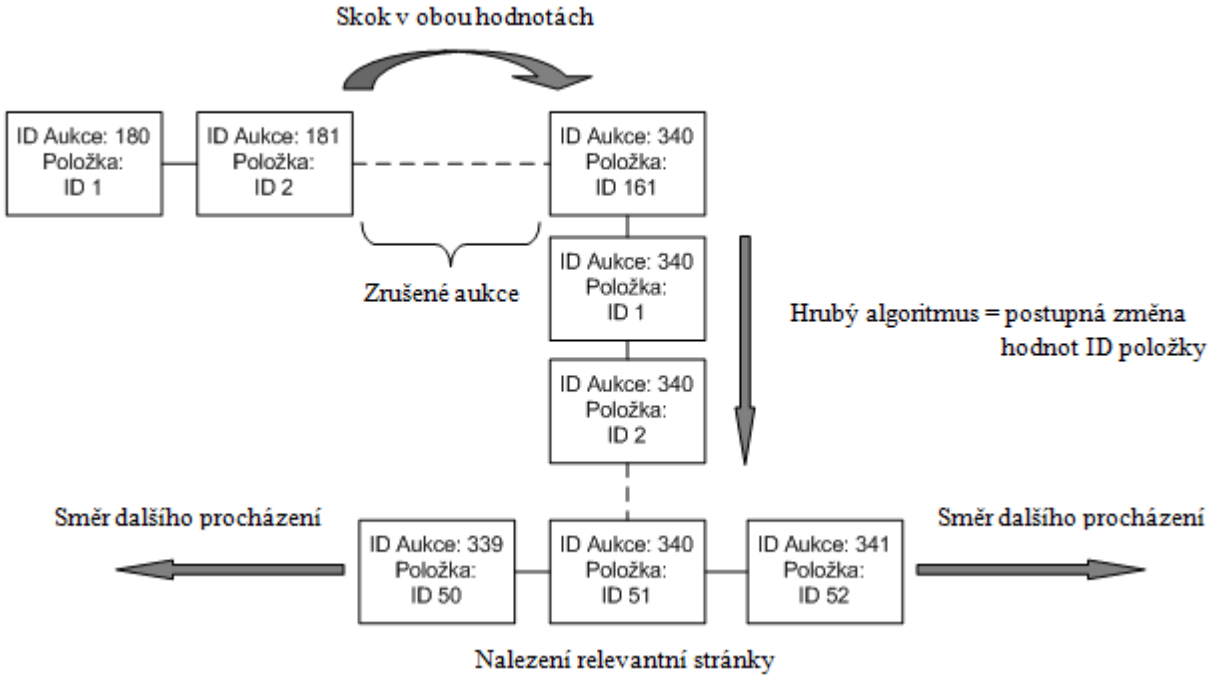
2b) Sekvence ID s dvojím číslováním

Dvojice číslování odděluje jednoznačné určení aukce a konkrétní položky v dané aukci (obrázek 13). V tomto případě se procházejí všechny záznamy jedné aukce a po dosažení poslední položky se přechází na novou hodnotu aukce.



Obrázek 13: Sekvence s dvojím číslováním

K největším komplikacím patří vynechávání sekvencí s náhodným pořadím v rámci dvojitého číslování. Eliminace tohoto problému souvisí se skoky v ID, které jsou popsány výše a zkoušením generování URL pomocí hrubé síly (obrázek 14) spojené se skoky a rozdvojováním posunů směrem dolů i nahoru v zadaných hodnotách.



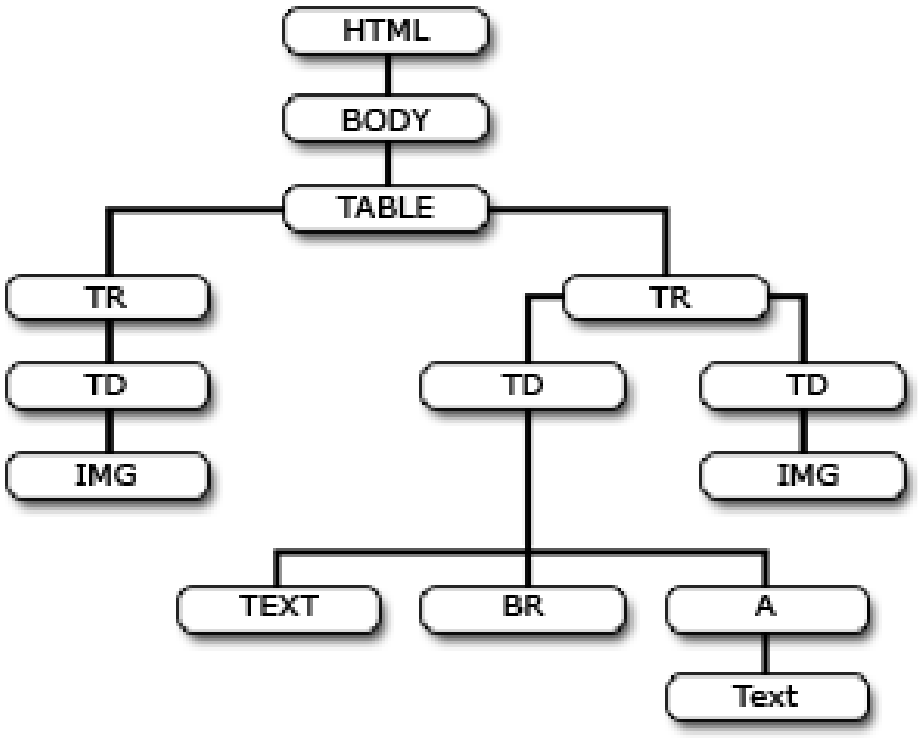
Obrázek 14: Sekvence s hrubým algoritmem a skoky v ID

Takto implementovaný způsob procházení sekvencemi dvojic číselných hodnot získává největší objem dostupných dat vyhovujících podmínkám uživatele. Další výhodou tohoto řešení je schopnost přesunutí se do další aukce s využitím skoku v obou hodnotách.

2.4.3 Zpracování webové stránky

Webová stránka zobrazuje informace poskytované v rámci celosvětové sítě. Informace jsou prezentovány v podobě hypertextu, který je vytvořen použitím značek HTML nebo XHTML. Stránky se skládají z textu, multimediálních dat jako jsou obrázky, videa, zvuky a odkazů, které umožňují přechod na další webové stránky. Strukturu těchto stránek bychom si mohli představit jako stromovou hierarchii (obrázek 15).

Výhodou řešení tohoto systému je, že dokáže podle přesně stanovených pravidel získávat pouze ty informace, které uživatel požaduje. Program je tedy vytvořen za účelem použití na různorodých zdrojových strukturách s využitím znalosti HTML a jeho aplikování.

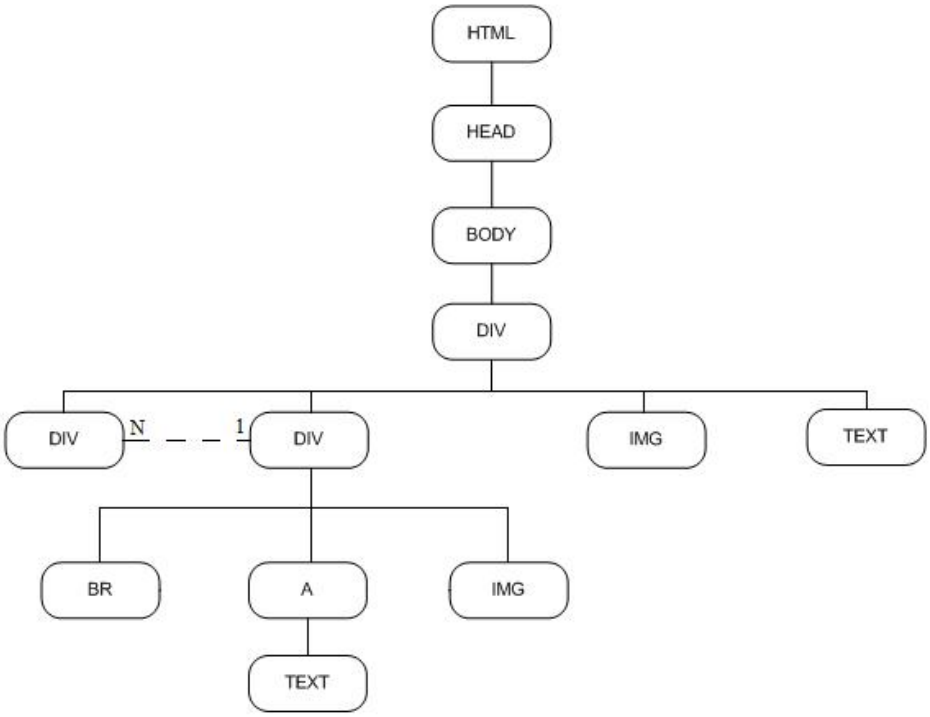


Obrázek 15: HTML strom

Přínosem této práce je, že systém nabízí uložení nastavení systému do šablony, kterou je možné znovu použít na internetových stránkách stejného typu pro doplnění již stažené databáze nebo k provedení její aktualizace. Šablona slouží jako předloha, podle které se budou internetové stránky zpracovávat.

Tento vzor je vytvořen na základě struktury obecných internetových stránek, která se vždy v cílovém hypertextovém dokumentu vyskytuje a upřesňuje, jak se má daná část dokumentu zpracovat. Práce s těmito vzory a šablonami je velice přínosná, protože usnadňují, zrychlují a zefektivňují výsledky potřeb uživatele. Vzhledem k možnostem snadné úpravy šablony a předlohy způsobu práce s internetovou stránkou, je systém schopen reagovat na změnu zdrojového kódu hypertextového dokumentu.

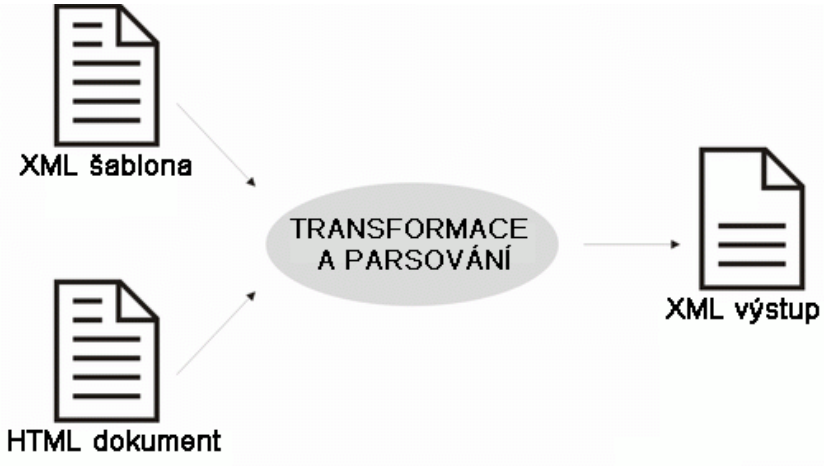
Modernější struktury webových stránek využívají značku DIV (obrázek 16), která je blokovým elementem. Tato značka nemá speciální význam a také se nijak neformátuje. Je důležitá zejména v kombinaci s kaskádovými styly a může obsahovat různé množství značek, které nesou svůj význam, liší se svou funkcí a vzhledem.



Obrázek 16: Struktura značky DIV

Implementací a provedením šablon, které kopírují strukturu internetových stránek, znamená, že systém není nastaven na fixní předem určenou množinu obecných hypertextových dokumentů. V této práci je zajištěna flexibilita a použitelnost na odlišná zdrojová data, která jsou v cílové podobě seskupena v jednotné formě.

Vstupními daty programu je šablona pro zpracování a internetová stránka, na kterou chceme konkrétní vzor použít. Výstupem je jednotný soubor typu XML. Systém se postará o aplikovaní šablony a převedení dat na výstup. Za běhu tedy proběhne proces (obrázek 17), který vše zpracuje a ponechá pouze korektní data.



Obrázek 17: Transformace dat

3. Funkce vytvářeného řešení

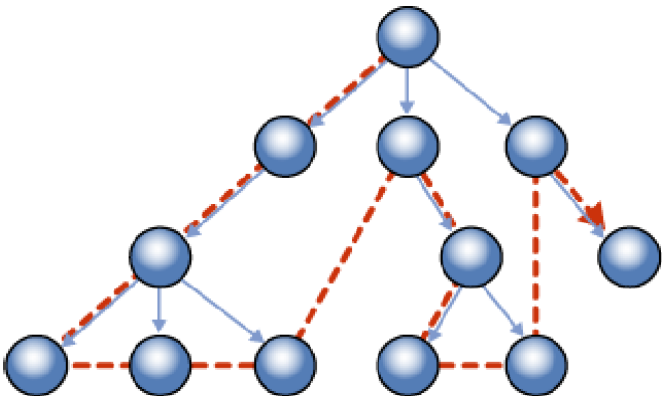
Pro snadnou a přehlednou orientaci práce v programu bylo zvoleno příjemné grafické rozhraní. Celý program je řešen ve třech různých barevných provedeních, které si může uživatel zvolit. K dalšímu usnadnění velice přispělo rozdělení samostatných nastavení do přehledných záložek, kdy je osoba pracující se systémem informována o průběhu nastavení programu názorným diagramem. Detailní popis použití grafického rozhraní programu je popsán v příloze B „Uživatelský manuál“. Tento vzhled byl docílen použitím trial verze komponenty Infragistics [5].

K přednostem tohoto řešení patří, že nabízí možnost zvolit si procházení internetových stránek pomocí hledání nových odkazů nebo generováním URL. Při vybrání jedné z nabízených možností je uživatel plynule naváděn k zadání dalších údajů potřebných pro vykonání požadovaných operací.

3.1 Hypertextové odkazy

Funkce procházení hypertextových odkazů pracuje na způsobu webového robota, který hledá nové dostupné odkazy splňující zadaná kritéria. Pro zajištění konečnosti algoritmu programu je nastaveno, že systém bude zjišťovat informace pouze na počáteční doméně, kterou zadal uživatel. Prostředky k detailnějšímu upřesnění umožní definovat užší oblast internetových stránek, které se mají zpracovat. K této problematice má osoba pracující se systémem nástroj, kdy přesně stanoví text, který musí požadovaný odkaz obsahovat.

K zajištění zanořování systému v rámci internetových stránek se nastavuje hloubka, kterou program prochází. Výběr hloubky je reprezentován relativní hodnotou k počáteční URL. Algoritmus prohledávání do hloubky (obrázek 18) pracuje tak, že vždy expanduje prvního následovníka každého vrcholu, pokud jej ještě nenavštívil. V situaci, kdy narazí na vrchol, z nějž už nelze dále pokračovat, vrací se zpět backtrackingem [6].



Obrázek 18: Algoritmus procházení do hloubky

Z dosaženého vrcholu se nedá dále pokračovat, pokud nemá žádné následníky nebo již byli všichni navštíveni. V tomto případě se použije backtracking neboli zpětné vyhledávání, kdy se algoritmus vrací k předchozímu vrcholu a provádí kontrolu možnosti dalšího pokračování.

Na internetových stránkách se nachází různé druhy odkazů. Většina standardních hypertextových odkazů je reprezentována HTML značkami, které definují jejich přesnou funkci. Jiné formy odkazů na další dokumenty jsou obsaženy v externích stylech, skriptech nebo javovských souborech. Přehled odkazů, které systém umí zpracovat, naleznete v tabulce 2.

Typ odkazu	Realizace
HTML tag <A>	Ano
HTML tag <FRAME>	Ano
HTML tag <IFRAME>	Ano
HTML tag <AREA>	Ano
HTML tag <IMG>	Ano
Kaskádový styl (.css)	Ne
Flash (.swf)	Ne
Javascript (.js)	Ne
Java soubor (.class)	Ne

Tabulka 2: Podporované typy odkazů

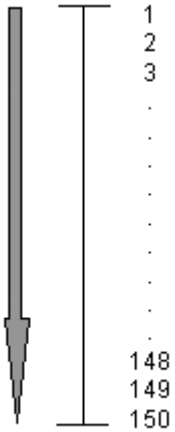
3.2 Generování URL

V případě, že se zpracovávají stránky, které jsou systematicky značeny, je efektivnější využít metodu generování odkazů. Systém nabízí tři varianty pro uspořádané vytváření URL. Všechny nabízené možnosti pracují s číselnými řetězci.

3.2.1 Zpracování jednoho parametru

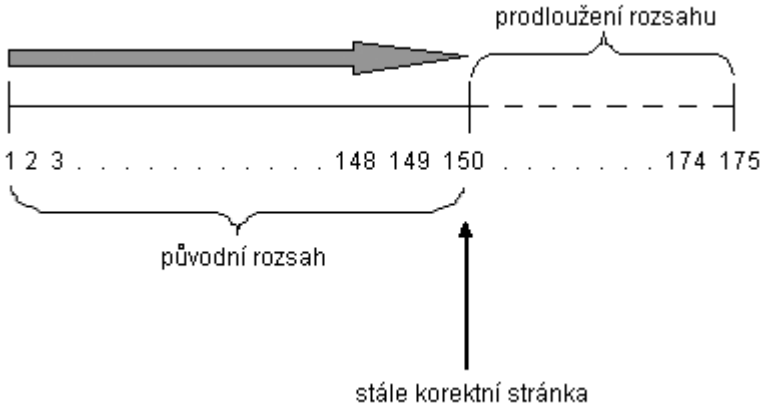
Uživatel ohraničí část textu v zadané URL speciálními značkami a zadá rozsah čísel pro něž chce, aby byly dosazeny do počátečního odkazu. Řešení je realizováno cyklem (obrázek 19), který postupně vytvořené odkazy kontroluje a webovou stránku, která je serverem vrácena, parsuje nebo zahazuje a přechází k dalšímu kroku cyklu.

V tuto chvíli systém generuje neupravenou posloupnost číselných hodnot a s každou dosazenou hodnotou do počáteční URL kontroluje relevantnost vrácených dat ze serveru.



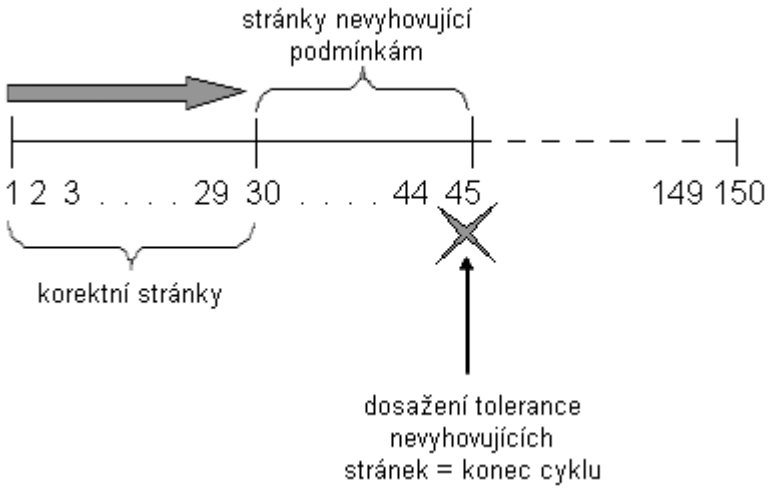
Obrázek 19: Procházení jedním cyklem

Pro lepší využití tohoto algoritmu lze definovat číselnou hodnotu, o kterou se zadaný rozsah rozšíří (obrázek 20) po skončení celého cyklu v případě, že poslední vygenerovaná stránka odpovídala kritériím pro parsování. Tato možnost je implementována pro situaci, kdy uživatel nezná přesný obsah dat a chce získat největší dostupné množství korektních informací.



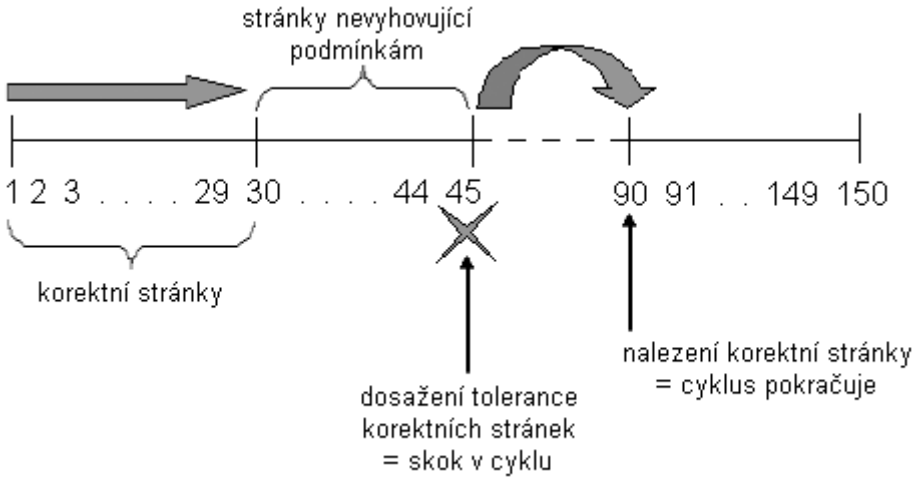
Obrázek 20: Zvýšení horní hranice rozsahu

K rozšířenosti tohoto cyklu také přispívá možnost předčasného ukončení (obrázek 21), pokud generované odkazy nevracejí potřebné hodnoty. Využitím této funkce se ušetří velké množství času v situaci, kdy by se měly vytvářet a zkoušet pouze stránky, které nesplňují zadané podmínky. V tomto případě, pokud webový server nevrací žádnou odezvu na odeslané požadavky, tedy webové stránky vůbec neexistují nebo vrací pouze stránky, které obsahují nerelevantní hodnoty. Koncový uživatel zadá hodnotu tolerance, a pokud po sobě jdoucí počet internetových stránek nevyhovujících kritériím pro další zpracování, cyklus končí.



Obrázek 21: Předčasné ukončení cyklu

Aby se předešlo předčasnému ukončení algoritmu v případě, že v zadaném rozsahu pro generování odkazů je posloupnost, která vrací nepotřebná data a za tímto úsekem se opět nachází internetové stránky splňující požadavky uživatele, tak lze nadefinovat číselný posun (obrázek 22). Určený posun se provádí, pokud je dosažena tolerance korektnosti stránek a skok v cyklu nepřesáhne aktuální maximální hodnotu rozsahu. V situaci, kdy po skoku stále nejsou nalezeny korektní stránky, se provádí další skok.



Obrázek 22: Skoky v cyklu

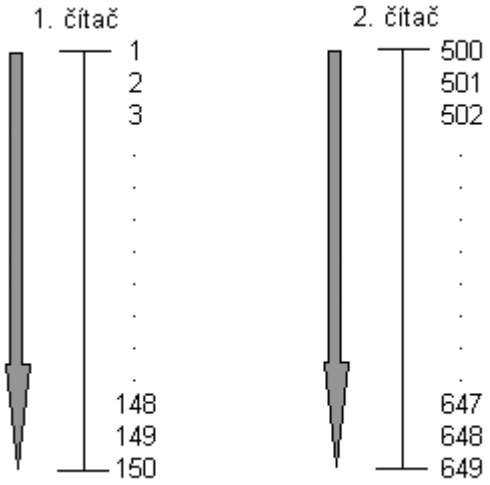
V případě, že je po skoku nalezena stránka, která obsahuje relevantní hodnoty, cyklus se rozděluje do dvou fází (obrázek 23). Prvně se cyklus vrací zpět k poslední známé hodnotě od provedeného posunu, protože nevíme, kde se sekvence opět navázala. Teprve poté čítač postupuje směrem vpřed. Touto implementací je docíleno získání maximálního obsahu dat.

```
13:16:38: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=376&AucID=362&Lot=3
13:16:41: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=377&AucID=362&Lot=4
13:16:43: Stránka se nebude parsovat: http://www.coinarchives.com/a/lotviewer.php?LotID=378&AucID=362&Lot=5
13:16:43: Stránka se nebude parsovat: http://www.coinarchives.com/a/lotviewer.php?LotID=379&AucID=362&Lot=6
13:16:43: Stránka se nebude parsovat: http://www.coinarchives.com/a/lotviewer.php?LotID=380&AucID=362&Lot=7
13:16:43: Skočil jsem na link: http://www.coinarchives.com/a/lotviewer.php?LotID=410&AucID=362&Lot=37
13:16:44: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=410&AucID=362&Lot=37
13:16:47: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=409&AucID=362&Lot=36
13:16:50: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=408&AucID=362&Lot=35
13:16:53: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=407&AucID=362&Lot=34
13:16:56: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=406&AucID=362&Lot=33
13:16:59: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=405&AucID=362&Lot=32
13:17:02: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=404&AucID=362&Lot=31
13:17:05: Stránka se nebude parsovat: http://www.coinarchives.com/a/lotviewer.php?LotID=403&AucID=362&Lot=30
13:17:05: Stránka se nebude parsovat: http://www.coinarchives.com/a/lotviewer.php?LotID=402&AucID=362&Lot=29
13:17:05: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=401&AucID=362&Lot=28
13:17:08: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=400&AucID=362&Lot=27
13:17:10: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=399&AucID=362&Lot=26
13:17:13: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=398&AucID=362&Lot=25
13:17:16: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=397&AucID=362&Lot=24
13:17:19: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=396&AucID=362&Lot=23
13:17:22: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=395&AucID=362&Lot=22
13:17:25: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=394&AucID=362&Lot=21
13:17:28: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=393&AucID=362&Lot=20
13:17:31: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=392&AucID=362&Lot=19
13:17:34: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=391&AucID=362&Lot=18
13:17:37: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=390&AucID=362&Lot=17
13:17:40: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=389&AucID=362&Lot=16
13:17:43: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=388&AucID=362&Lot=15
13:17:46: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=387&AucID=362&Lot=14
13:17:49: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=386&AucID=362&Lot=13
13:17:52: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=385&AucID=362&Lot=12
13:17:55: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=384&AucID=362&Lot=11
13:17:57: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=383&AucID=362&Lot=10
13:18:00: Stránka se nebude parsovat: http://www.coinarchives.com/a/lotviewer.php?LotID=382&AucID=362&Lot=9
13:18:01: Stránka se nebude parsovat: http://www.coinarchives.com/a/lotviewer.php?LotID=381&AucID=362&Lot=8
13:18:01: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=411&AucID=362&Lot=38
13:18:05: Budu parsovat stránku: http://www.coinarchives.com/a/lotviewer.php?LotID=412&AucID=362&Lot=39
```

Obrázek 23: Výpis logu generování stránek

3.2.2 Současné zpracování dvou parametrů

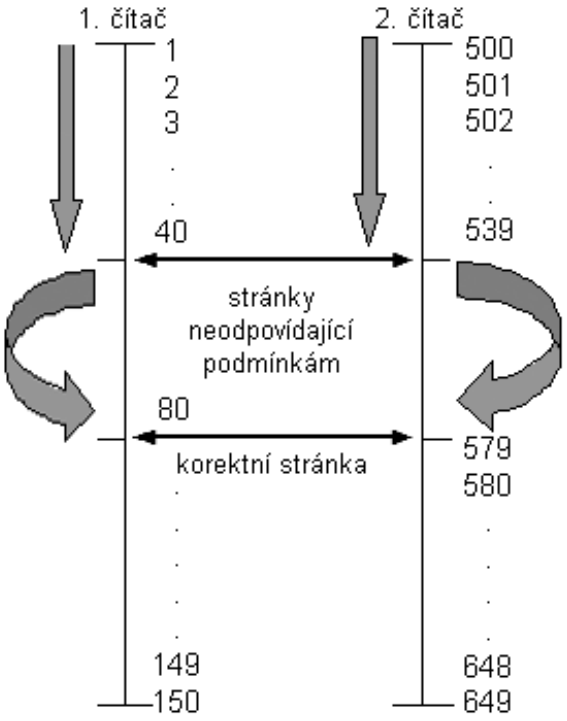
Při využití funkce současného zpracování dvou parametrů dochází ke změně odlišných čítačů (obrázek 24). Vyznačením místa v počátečním odkazu je určena oblast, do které se budou dosazovat číselné hodnoty, ze dvou rozlišných rozsahů.



Obrázek 24: Současná změna parametrů

Proces generování nových odkazů se ukončí po dosažení horního limitu jednoho z cyklů. Číselné hodnoty dosazované do odkazů jsou zvyšovány současně. V situaci, kdy jsou získávány internetové stránky, které nesplňují uživatelem definovaná pravidla, se první čítač zastaví na aktuální hodnotě a druhý cyklus prochází celý rozsah, dokud není vrácena vyhovující stránka.

Pokud ani poté nejsou dosaženy potřebné výsledky, tak se provádí skok v obou cyklech zároveň (obrázek 25) a pokračuje se v generování odkazů od místa posunu čítačů.

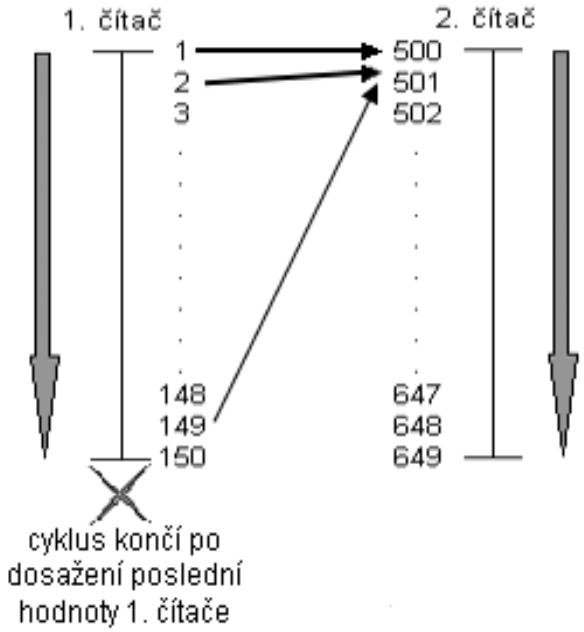


Obrázek 25: Současný posun v čítačích

3.2.3 Postupné zpracování dvou parametrů

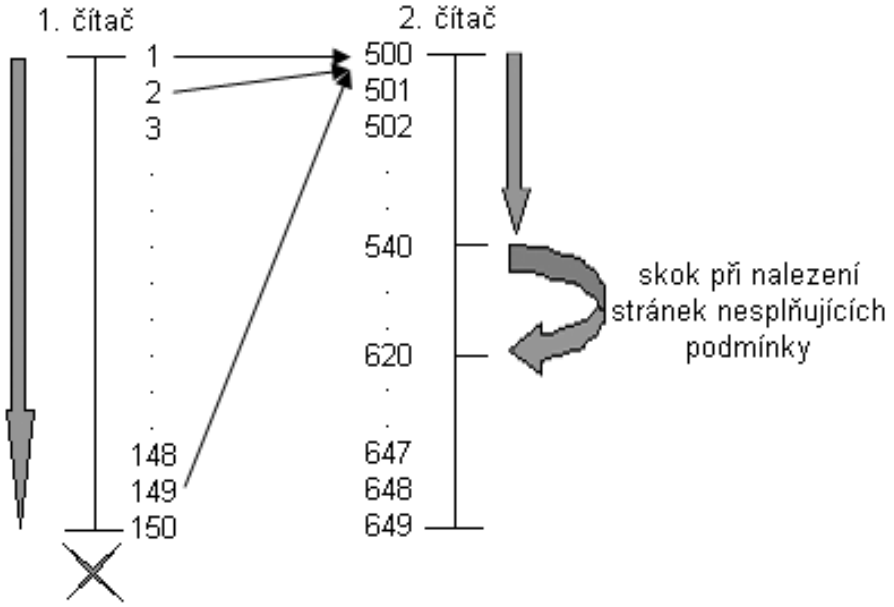
V případě parametrů, které se zpracovávají postupně, dochází k rozlišné změně obou čítačů (obrázek 26). Do označené části počáteční URL se dosadí číselná hodnota z rozsahu pro první čítač a v takto upravené URL se do druhé označené části vepisuje celý rozsah dalšího cyklu. Po skončení rozsahu druhého čítače se hodnota prvního čítače zvýší o hodnotu 1 a opět se dosazuje celý rozsah druhého cyklu.

Tento algoritmus je poté dále rozšířen o funkce, které ho značně zefektivňují. Následující rozšíření jsou popsána níže v této kapitole.



Obrázek 26: Postupná změna čítačů

Systém pokračuje v generování odkazů, dokud se první cyklus nedostane na konec rozsahu. V této situaci dochází k postupnému inkrementu dvou cyklů. Pro zefektivnění algoritmu je navržen způsob provádění posunu (obrázek 27) v hodnotách druhého čítače.

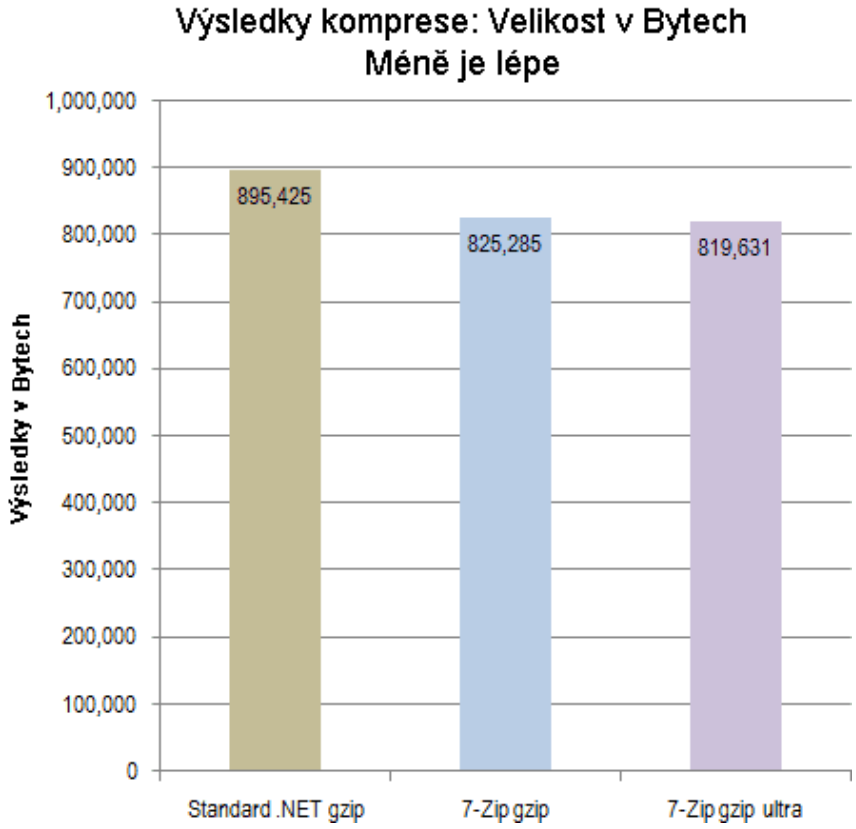


Obrázek 27: Posun v druhém čítači

3.3 Archivace souborů

Systém je vybaven archivací všech stažených souborů. Do této oblasti patří, jak vytvořené výstupní soubory, tak všechny stažené obrázky a vytvořené složky. V archivovaném souboru se vytvoří přesná kopie dat, které byly na disk uloženy. Výhodou této realizace je úspora místa na disku. Program nabízí možnost ponechání pouze cílového archívu a smazání dat, které jsou do archívu přidány.

Z důvodu velikosti komprese (obrázek 28) byl zvolen způsob archivace pomocí využití programu 7-Zip. V detailnějším nastavení tohoto algoritmu je vybrána metoda, pro největší úsporu místa na disku.

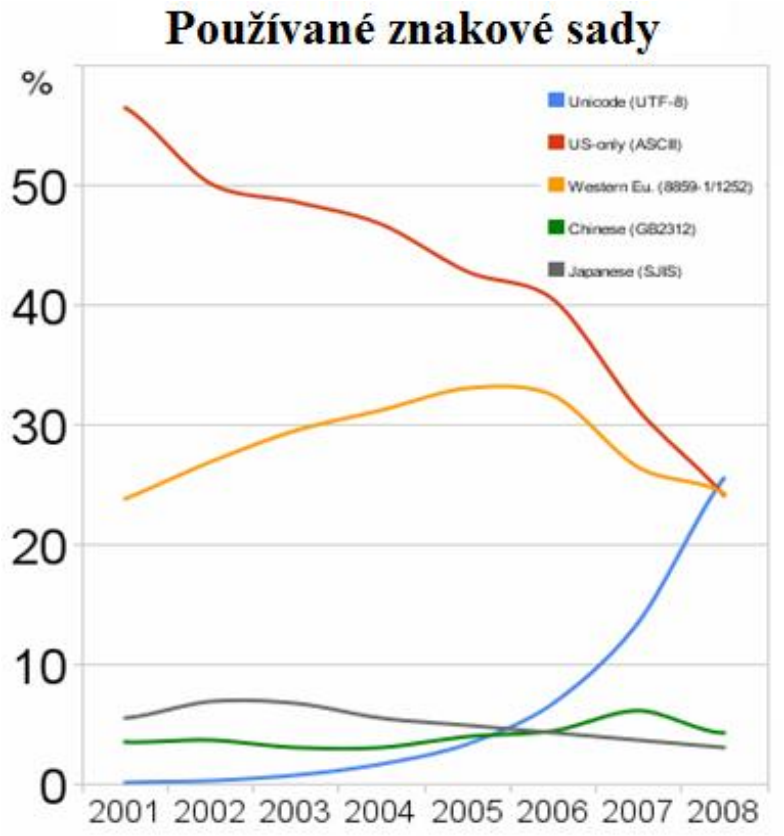


Obrázek 28: Porovnání komprese algoritmů

3.4 Kontrola kódování

V průběhu načítání internetové stránky systém zjišťuje znakovou sadu, ve které je obsah zapsán. V případě, že program získá nastavení kódování, tak je pro převod stránky použito právě aktuální kódování. Pokud ovšem není zjištěna vstupní znaková sada, systém použije jako výchozí kódování Windows-1250. Všechny výstupní soubory se poté převádějí na znakovou sadu UTF-8 [7].

UTF-8 je zkratka pro UCS Transformation Format. Je to způsob kódování řetězců znaků Unicode/UCS do sekvencí bajtů. Vývoj použití znakové sady UTF-8 (obrázek 29) v rámci webu stále vzrůstá a v současné době patří k nejpoužívanější.



Obrázek 29: Používané znakové sady na internetových stránkách

Z důvodu menšího využití pásma rychlosti pro připojení k Internetu a neblokování linky pro ostatní uživatele, lze v systému vybrat možnost kontroly kódování pouze na první internetové stránce a aplikování tohoto nastavení na všechny následující stránky.

3.5 Úprava výsledných souborů

Po dokončení metod, které zajišťují stahování obsahu a jeho uložení do koncového souboru, se provádí kontrola korektnosti. K hlavní činnosti patří uzavření všech elementů a vytvoření šablony pro grafické znázornění. Šablona zajišťuje a definuje styl, jakým se mají jednotlivé elementy zobrazovat. Výhodou řešení je nezávislost na obsahu výsledného souboru. Šablona se vytváří na základě struktury vytvořeného souboru, a tedy reaguje na rozlišná zadání a podmínky uživatele.

3.5.1 Třídění souborů

K třízení výsledných dat je vytvořena šablona typu XSL. Takto vytvořená šablona je generován na základě struktury a obsahu cílového souboru. Samotný proces je poté proveden pomocí transformace.

Transformace XSLT

Transformace XSLT slouží k převodům zdrojových dat ve formátu XML do libovolného požadovaného formátu, nejčastěji HTML, jiného XML nebo jiných datových struktur.

XSLT je transformace, která se provádí pomocí procesoru XSLT. Procesorem je uvažován program podporující tuto transformaci. Procesor XSLT může být napsán v libovolném programovacím jazyce nebo využít knihovny XSLT [8] daného jazyka.

K provedení transformace jsou potřeba dva soubory. První soubor obsahuje zdrojová data, která budou transformována. Struktura tohoto souboru vyjma obecných vlastností XML není blíže specifikována. Druhý soubor obsahuje vzorec pro transformaci a musí být napsán v jazyce XSL (obrázek 30).

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  - <xsl:template match="/">
    - <html>
      - <body>
          <h2>My CD Collection</h2>
        - <table border="1">
          - <tr bgcolor="#9acd32">
              <th>Title</th>
              <th>Artist</th>
            </tr>
          - <xsl:for-each select="catalog/cd">
            - <tr>
              - <td>
                  <xsl:value-of select="title" />
                </td>
              - <xsl:choose>
                - <xsl:when test="price > 10">
                  - <td bgcolor="#ff00ff">
                      <xsl:value-of select="artist" />
                    </td>
                  </xsl:when>
                - <xsl:when test="price > 9">
                  - <td bgcolor="#cccccc">
                      <xsl:value-of select="artist" />
                    </td>
                  </xsl:when>
                - <xsl:otherwise>
                  - <td>
                      <xsl:value-of select="artist" />
                    </td>
                  </xsl:otherwise>
                </xsl:choose>
              </tr>
            </xsl:for-each>
          </table>
        </body>
      </html>
    </xsl:template>
</xsl:stylesheet>
```

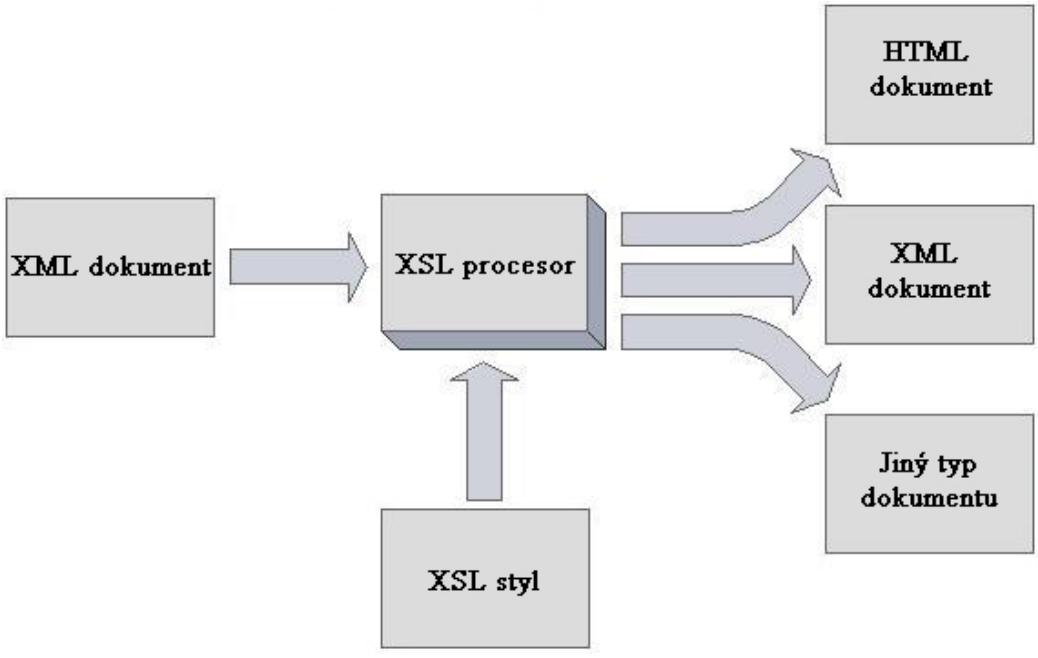
Obrázek 30: Vzorový soubor XSL pro transformaci

Zdrojová data pro transformaci mohou obsahovat libovolné znakové sady, nejen angličtinu, ale také znaky s diakritikou nebo znaky asijských písem, popřípadě jejich kombinace. V zdrojovém dokumentu však musí být v záhlaví označeno kódování znaků. V poslední době se velice využívá a je doporučeno užívat kódování UTF-8 pro správnou funkci jiných znaků než ASCII. Většina operačních systémů, programovacích jazyků a editorů je již podporou kódování UTF-8 vybavena.

Soubory pro XSLT jsou v textovém formátu, který je čitelný bez žádného zvláštního editoru. Díky tomuto principu je i možné tyto soubory snadno generovat pomocí počítačových programů. Příkladem může být získání dat z databáze či jiných datových struktur, jejich konverze do XML, která je velmi snadná, a následovná aplikace XSLT například pro převod do HTML.

Procesory nebo též programy pro provedení vlastní transformace existují pro všechny známé platformy používané na PC nebo na serverech. V mnoha programovacích jazycích jsou dnes již implementovány knihovny pro XSLT.

Smyslem XSLT (obrázek 31) je na základě zdrojového souboru a šablony vygenerovat jiný, třetí dokument nebo obecně soubor. Struktura tohoto výstupu XSLT není definována přímo standardem a je závislá na procesoru XSLT. Nejčastěji se používá výstup do HTML nebo XML, případně prostý textový formát, označovaný též TXT. Dalšími velmi známými výstupy jsou formáty PDF a RTF. Zcela pochopitelně to však mohou být i libovolné jiné soubory nebo formáty dat.



Obrázek 31: XSL transformace

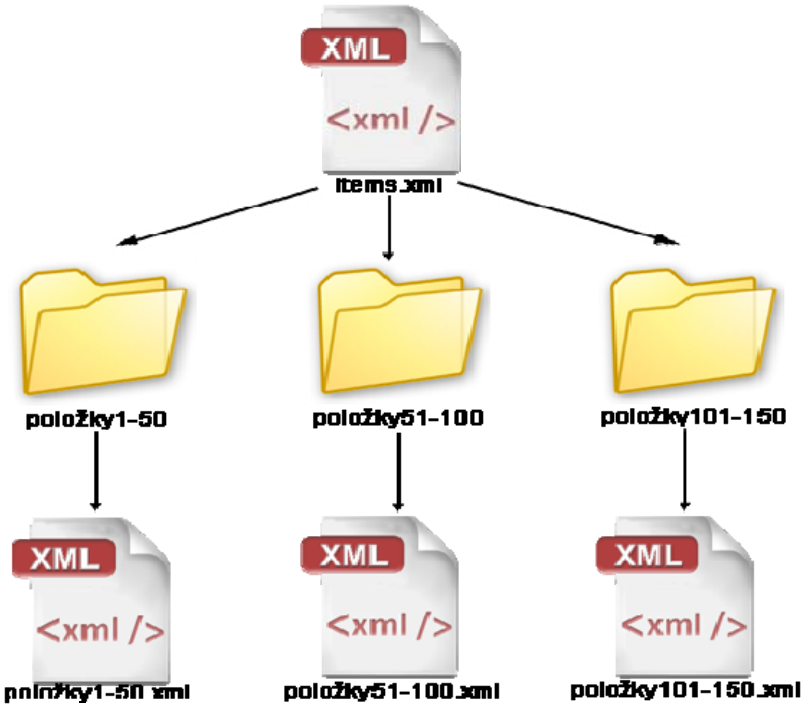
3.5.2 Rozdělení souborů

Systém dokáže rozdělit výsledný soubor obsahující všechny položky na jednotlivé dílčí části. Vznikne několik souborů typu XML, které obsahují přesný rozsah položek s jednotným pojmenováním nadefinovaným uživatelem.

Tento způsob byl realizován z důvodu pozdějšího zatřízení nových stažených dat do příslušných výstupních souborů. Pozdější práce s takto fyzicky rozdělenými úseky dat je mnohem efektivnější a rychlejší z důvodu menšího vytížení operační paměti při načítání souboru. Pro uživatele je druhou významnou výhodou snadnější orientace a přehlednost v strukturovaných položkách.

Při využití generování URL se rozdělená struktura vytváří již za běhu a jednotlivé položky se zapisují do již předem připravených souborů. Této možnosti lze využít za pomoci čítačů, kdy jejich hodnota případně dvojice hodnot reprezentuje klíč položky.

V případě procházení hypertextových odkazů je vytvořen dočasný soubor, pro uložení dat. Po dokončeném stahování a zkompletování dočasného souboru probíhá postupné rozdělení (obrázek 32) do jednotlivých částí. Rozdělení je realizováno podle elementu položky, který obsahuje číselnou hodnotu. Při dosažení poslední položky je dočasný soubor smazán a všechna nově vytvořená data jsou zkontrolována. V průběhu rozdělování jsou také přesouvány potřebné stažené obrázky.



Obrázek 32: Rozdělení souborů

3.6 Načtení dat ze souboru

Z již známých stažených dat lze vyexportovat seznam čísel nebo dvojice čísel, která jednoznačně reprezentují položku. Pro načtení čísel je využit textový soubor, ve kterém jsou všechna data uložena. Způsob zpracování zadaných číselných hodnot je používán při generování URL.

Hodnotám lze přiřadit vlastnost, zda se mají přesně tyto data zpracovat a generovat URL nebo jestli se v prováděných cyklech mají vynechat. V případě práce se vstupními čísly jsou postupně dosazovány (obrázek 33) do jednotlivých zvolených čítačů.

Obrázek 33: Zpracování dat ze souboru

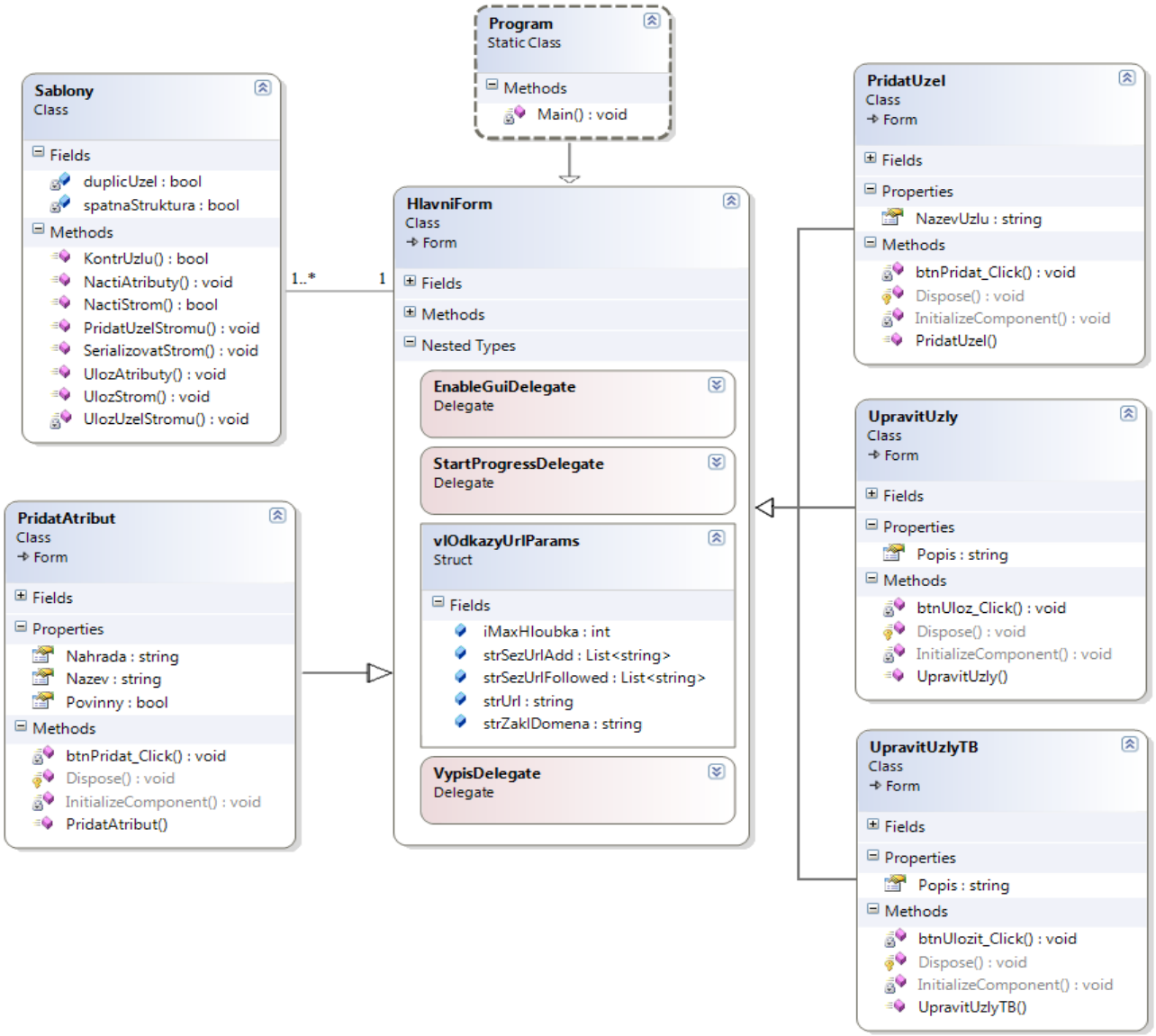
4. Analýza systému

V této kapitole je popsán postup realizace zadaného problému, který si více přiblížíme v následujících sekcích:

- třídní diagram
- popis tříd
- testovaná data

4.1 Třídní diagram

Vztahy mezi jednotlivými třídami (obrázek 34) zobrazují základní strukturu.



Obrázek 34: Třídní diagram

Třídní diagram reprezentuje návrh programu a propojení jeho jednotlivých tříd. Zaznamenány jsou také metody a proměnné, které daná třída využívá.

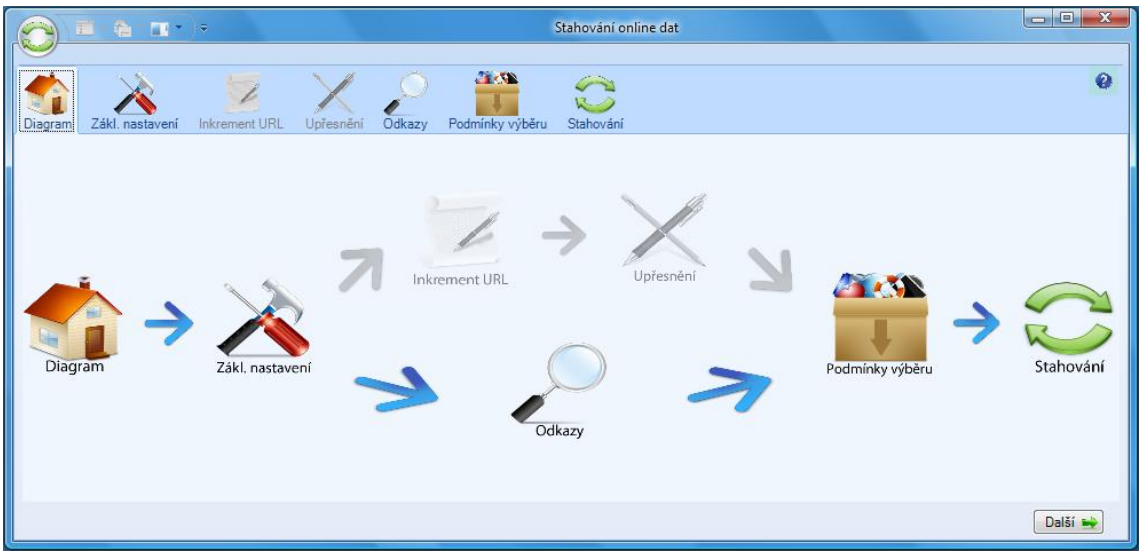
4.2 Popis tříd

Program:

Třída obsahující metodu „Main“, která je vstupním bodem programu a spouští celý systém. Tato metoda zavolá zobrazení hlavní nabídky.

HlavniForm:

Formulář, který je reprezentován touto třídou, poskytuje uživateli hlavní grafické rozhraní a zpřístupňuje konkrétní nastavení. Design a styl zobrazení je přizpůsoben k usnadnění práce. Pro názornou grafickou reprezentaci a orientaci v systému byl navržen diagram (obrázek 35). Diagram neslouží pouze pro ukázání průběhu nastavení, ale při stisknutí tlačítka nad požadovaným obrázkem, se přesune ke specifické části programu. V případě, že daná část programu není zpřístupněna, je grafická ikona zastupující konkrétní nastavení zašedlá a možnost provést stisknutí nad touto ikonu je zablokován.



Obrázek 35: Záložka Diagram

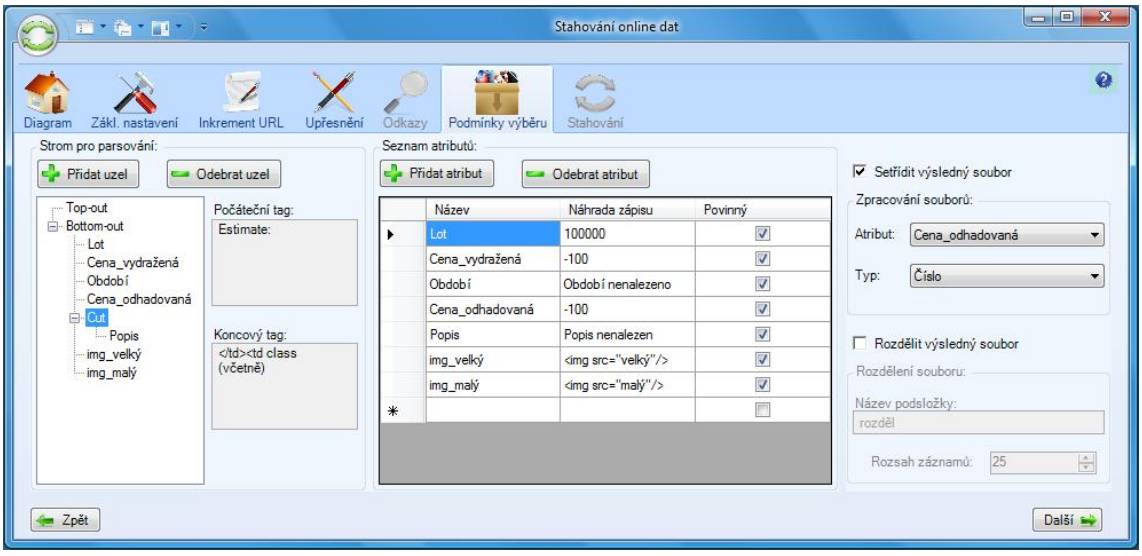
Tato třída také zajišťuje výběr způsobu hledání požadovaných dat. Na výběr jsou možnosti procházení hypertextových odkazů nebo generování URL. Rozdíl ve výběru je naznačen i v případě diagramu. Pokud je vybrána metoda procházení hypertextových odkazů, tak poté se ihned nastavují obecná pravidla pro zpracování stránky. V situaci, kdy je nastaveno generování URL, tak se zpřístupní nová položka, ve které je možnost vybrat načtení vstupních dat ze souboru nebo podmínka pro stahování obrázků.

Provedení výběru jedné z možností se nachází na záložce „Základní nastavení“ (obrázek 36), která obsahuje všechny základní informace potřebné pro další nastavení.

Obrázek 36: Záložka Základní nastavení

Pokud uživatel zvolil metodu procházení hypertextových odkazů, tak se využívá nadefinované struktury „vlOdkazyUrlParams“, která slouží jako jádro pro tento algoritmus. Struktura si udržuje informace o maximální hloubce, do které se může zanořit, seznamu již navštívených adres, seznamu následovaných adres, aktuální adresu URL a základní doménu, ze které celý algoritmus začínal.

Po nastavení pro jednotlivé mechanismy procházení internetových stránek, se obě metody setkávají v záložce „Podmínky výběru“ (obrázek 37), která umožňuje uživateli zadat obecné požadavky k zpracování stránek. K využití této možnosti je připraven strom pro zadání struktury.



Obrázek 37: Záložka Podmínky výběru

Tabulka zde slouží jako zdroj názvů pro jednotlivé uzly stromu. Vytvořeným uzlům stromu jsou přidány počáteční a koncové značky, podle kterých se program orientuje na internetové stránce.

Sablony:

Třída byla vytvořena za účelem usnadnění práce a zrychlení používání programu. Zajišťuje serializaci a deserializaci stromu, tabulky atributů a celkového programu. Všechny vytvářené typy šablon jsou realizovány jako soubory typu XML.

V případě, že uživatel není schopen z nějakého důvodu dokončit nastavení, tak využije této možnosti a příště může pokračovat od posledního provedeného úkonu. K další výhodě patří, že lze aplikovat stejnou šablonu na jiný druh internetových stránek, pokud mají shodnou obecnou strukturu. V této situaci také uživatel nemusí nastavovat vše od začátku, ale pouze vybere požadovanou šablonu a aplikuje ji.

PridatAtribut:

Formulář reprezentovaný touto třídou má za úkol přidávat nové záznamy do tabulky s atributy. Během přidávání nového atributu se provádí validace zadaných dat. Kontroluje se správnost názvu atributu, jestli neobsahuje neplatný znak, který by poté mohl vést k nekorektně vytvořenému výslednému souboru nebo zda již zadaný atribut v tabulce neexistuje.

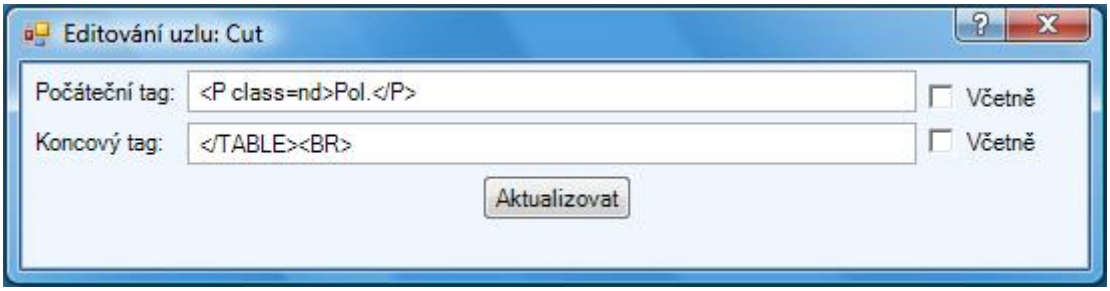
PridatUzel:

Formulář poskytuje uživateli možnost přidání nového uzlu do stromu a jeho přesné zatřízení. Třída se stará o identifikaci uzlu, na kterém byla událost vyvolána a zobrazí seznam dostupných názvů z tabulky atributů. Po vybrání příslušného názvu je vložen do stromu nový uzel, který se stává potomkem označeného uzlu před přidáním.

V průběhu přidávání nového uzlu je kontrolováno, zda již tento uzel strom neobsahuje. Pokud nastane situace duplicity názvu uzlu, je za uživatelem zvoleným jménem uzlu přidána číselná hodnota, která je s přibývajícím počtem shodných jmen navyšována. Nově vytvořený uzel obsahuje prázdné nastavení a je nutné jej vyplnit.

UpravitUzly, UpravitUzlyTB:

Formulář slouží k nastavení hodnot vybraného uzlu (obrázek 38). Tato možnost je vyvolána stisknutím pravého tlačítka nad požadovaným uzlem. K zadání je potřeba vyplnit počáteční a koncovou značku, podle které se bude program orientovat na zpracovávané stránce. Rozšířenou možností je, zda se má pracovat i se zadanou značkou nebo nikoliv.

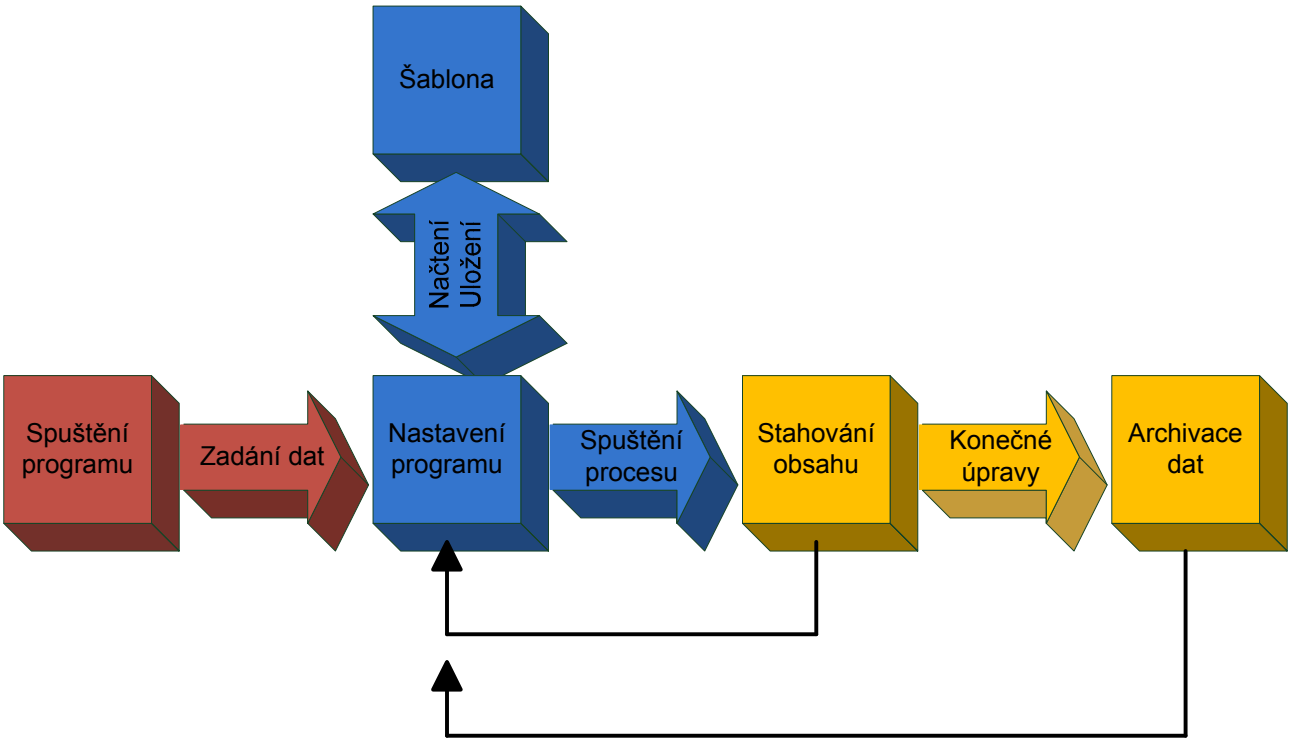


Obrázek 38: Nastavení hodnot uzlu

4.3 Blokový diagram

Diagram znázorňuje jednotlivé činnosti systému (obrázek 39), které lze provádět. Zobrazeny jsou také návaznosti mezi činnostmi. Práce s programem, kterou ovlivňuje uživatel je označena červenou a modrou barvou. V tuto chvíli je stále možné program ukončit, případně provést jiné požadované změny.

Po spuštění procesu již uživatel nemá možnost regulérně ovlivnit chod programu. Systém zpracovává zadaná kritéria a po dokončení všech operací a vrácení získaných dat, se program opět zpřístupní. Uživatel může reagovat na obdržené výsledky změnou požadavků nebo zadat kritéria pro jinou datovou strukturu a cyklus opakovat.

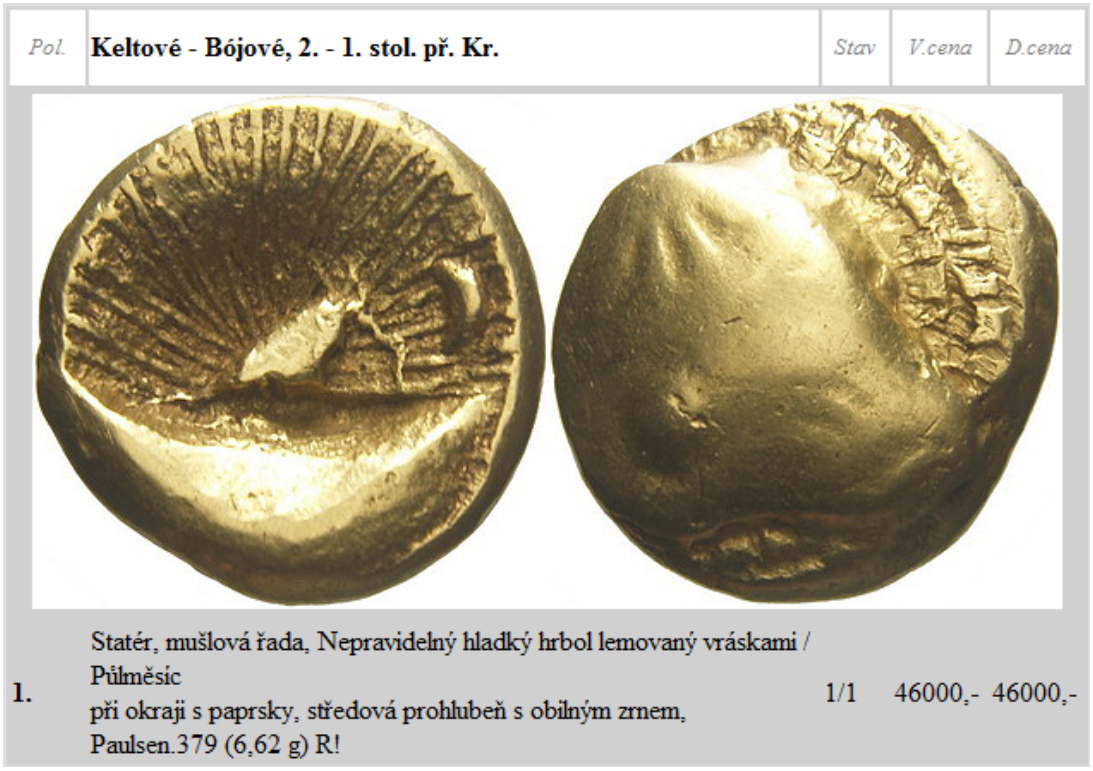


Obrázek 39: Blokový diagram systému

4.4 Testovaná data

Systém byl otestován na aukcích, které jsou specializovány v oboru numismatika. Byly provedeny testy, které ověřovaly celkovou funkčnost programu i dílčí úseky. Z důvodu rozlišnosti vstupních dat byly jako zdroje využity 4 různé online aukce, na kterých byla ověřena univerzálnost parsování a zpracování potřebných internetových stránek.

Pro demonstraci byla vybrána aukce AUREA Numismatika [9], kdy její struktura (obrázek 40) je tvořena více druhy mincí na jedné stránce. Šablona, která byla aplikována, získávala pouze potřebná data a ty poté byly uloženy do jednotné struktury. Pohyb na aukci se prováděl hledáním nových hypertextových odkazů s nastavením maximální hloubky zanoření.



Obrázek 40: Náhled položky na internetové stránce

Nastavení šablony a požadovaných dat:

- Období
- Lot
- Popis
- Kvalita
- Vstup_cena
- Konc_cena

Po uložení celkového obsahu na disk, proběhla kontrola výsledného souboru a byla vytvořena šablona XSL, která definuje vzhled výsledného souboru. Zobrazení dat (obrázek 41) v prohlížeči vypadá následovně a výsledný soubor obsahuje pouze požadované hodnoty.

Období: Keltové - Bójové, 2. - 1. stol. př. Kr.

http://www.aurea.cz/Katalog22/22a0001a.jpg
http://www.aurea.cz/Katalog22/22a0001r.jpg

Lot: 1
Popis: Statér, mušlová řada, Nepravidelný hladký hrbol lemovaný vráskami /
Půlměsícpři okraji s paprsky, středová prohlubeň s obilným zrnem, Paulsen.379 (6,62 g) R!
Kvalita: 1/1
Vstup_cena: 46000,-
Konc_cena: 46000,-

Obrázek 41: Náhled položky na disku

Všechny položky jsou tvořeny jednotnou strukturou, která je velice snadno přístupná. Práce s daty je jednoduchá pomocí jazyků XQuery nebo XPath. Výsledný soubor se v tuto chvíli stává databází, která v sobě uchovává pouze ty informace, které uživatel vyžaduje.

CoinArchives:

Tento typ aukce [10] je realizován způsobem, kdy na internetové stránce se nacházejí informace pouze o jedné minci. K procházení je využit algoritmus generování nových URL. Konkrétně definován současnou změnou dvou čítačů.

Otestována byla funkčnost stahování obrázků, které jsou zobrazeny jako odkaz (obrázek 42). Takto vložený obrázek reprezentuje miniaturu a přesměrovává uživatele na plnou velikost zobrazení (obrázek 36).

Obrázek 35: Miniatura obrázku simulující odkaz

Obrázek 42: Plná velikost obrázku

Po zpracování stránky jsou do souboru uložena data nesoucí informace o konkrétní minci a výsledná položka obsahuje miniaturu obrázku a zároveň jeho plnou velikost (obrázek 43).

K jednotlivým obrázkům jsou také přiřazeny odkazy, které reprezentují absolutní internetovou adresu. Hlavní účel těchto adres slouží ke kontrole, zda byly uloženy korektní obrázky nebo k jejich dodatečnému stažení, pokud by došlo k ztrátě dat.

Lot_ID: 255164
Číslo_lotu: 8026
Lot: 8026
Cena_vydražená: Unsold
Období: KELTISCHE MÜNZEN GALLIA AULERCI EBUROVICES
Cena_odhadovaná: 1000 EUR
Popis: AV-1/3 Stater, 2./1. Jahrhundert v. Chr.; 2.96 g. Stilisierter Kopf l.//Pferd r. mit Lenker, darunter Wolf. Delestrée/Tache 2395. R Kl. Randausbruch, sehr schönvorzüglich

http://imagedb.coinarchives.com/img/kunker/153/image08026.jpg
http://imagedb.coinarchives.com/img/kunker/153/thumb08026.jpg

Obrázek 43: Náhled uložených dat položky

Jiné testované aukce:

Do dalších testů byly zahrnuty internetové aukce Sixbid [11] a Stack's [12]. První zmiňovaná sloužila k otestování zbylých dvou algoritmů pro generování URL. K těmto algoritmům patří změna jednoho čítače a postupná změna dvou čítačů, kdy první čítač zastupuje číslo aukce a druhý hodnotu specifického záznamu v aukci.

Na internetové aukci Stack's bylo ověřeno zarovnání dosazovaného čísla na určitý počet znaků. Jako prázdné znaky se zvolila hodnota nula.

E-Shop:

Stahování obsahu E-shopu je opět prováděno na základě zadaných kritérií uživatelem. Internetové obchody k distribuci svých výrobků nejčastěji využívají možnost RSS kanálu. Vytvořený systém opět zvládá provedení parsování zdrojového souboru. V tomto případě je prováděna transformace dat ze struktury XML do jiné struktury XML, která odpovídá uživatelovým představám.

5. Provádění plánovaných aktivit

Kapitola popisuje principy řešení simulace chování reálného uživatele v prostředí internetových aukcí a provádění plánovaných aktivit. K těmto akcím patří možnost automatického přihlášení na registrované stránky a realizace samostatných příhozů.

V rámci dodržení základních pravidel a principů obecné aukce bývá využito funkce, která v případě příhozu provedeného v těsné blízkosti před ukončením řádného dražebního termínu, prodlouží dobu expirace dražby o určitý časový interval. Tato možnost nabízí uživatelům provést reakci na poslední událost.

Definováním automatického příhozu je snaha pomoci uživateli zabránit situaci, kdy je jeho nabídka převýšena příhozem na poslední chvíli. V aukčních systémech, které nejsou realizovány způsobem prodloužení expirace aukce, dochází k provádění příhozů jednotlivých uživatelů teprve v okamžiku, kdy se blíží konečný termín možnosti zadat svou nabídku.

V tomto případě se systém automatických příhozů snaží minimalizovat riziko a co nejvíce eliminovat možnost, která by vedla k převýšení částkou od jiného uživatele.

Příhozy nad požadovanými položkami by měly ověřovat datum a čas ukončení dražby. Mezi další vlastnosti, které by měly být splněny, patří nastavení maximální možné hodnoty, kterou jsme ochotni nabídnout a výše přihazované částky.

Problémy, které souvisí s touto realizací, jsou provedení přihlášení na stránkách, které využívají protokol HTTPS a pozdější udržení relace mezi uživatelem a webovým serverem za pomoci SID. K internetovým aukcím, které jsou takto spravovány, patří ze zahraničních eBay [13] a z domácích Aukro [14].

5.1 HTTPS

HTTPS je nadstavba síťového protokolu HTTP, která umožňuje zabezpečit spojení mezi webovým prohlížečem a webovým serverem před odposloucháváním, podvržením dat a umožňuje též ověřit identitu protistrany. HTTPS [15] používá protokol HTTP, přičemž přenášená data jsou šifrována pomocí SSL nebo TLS a standardní port na straně serveru je 443.

Protokol HTTPS využívá asymetrické šifrování. Obě strany si před zahájením komunikace vygenerují pár klíčů (privátní a veřejný). Při zahájení komunikace si vymění veřejné klíče, které by obě strany měly ověřit pomocí jiného komunikačního kanálu. Ověření může proběhnout kontrolou výtahu veřejného klíče u protistrany například pomocí telefonu.

Nebo lze použít princip přenosu důvěry, kdy nám protistrana předá veřejný klíč, který je digitálně podepsaný. Nejlépe certifikační autoritou, které důvěřujeme a jejíž veřejný klíč máme v důvěryhodném úložišti.

Zatímco samotné šifrování ochrání komunikaci před odposloucháváním, bez ověření autenticity veřejných klíčů jsou komunikující strany vystaveny riziku útoku MITM.

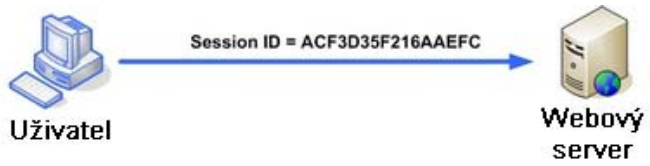
Při navázání spojení pomocí HTTPS je veškerá komunikace ihned od počátku šifrována pomocí SSL/TLS, není možné včas serveru sdělit, s jakým doménovým jménem chceme pracovat. Stejně jako s pomocí řádku Host: u protokolu HTTP. Proto pro HTTPS spojení standardně nelze vytvářet více virtuálních webových serverů na jediné IP adrese a portu, které jsou rozlišeny pouze doménovým jménem.

5.2 Relace – Session

Jako session se rozumí trvající síťové spojení mezi klientem a serverem, zahrnující výměnu většího množství paketů.

U protokolů jako je TELNET nebo FTP session odpovídá spojení na úrovni nižšího protokolu TCP. V případě použití protokolů, které žádnou podporu pro sessions nemají (UDP), nebo kde spojení typicky trvají velmi krátkou dobu (HTTP), jsou session udržovány přímo aplikačním programem, a k tomu nutné informace jsou vkládány do přenášených dat.

Typickým příkladem je použití HTTP cookie k uložení jednoznačného identifikátoru SID (obrázek 44), podle něhož pak server ve své paměti najde potřebné informace o přihlášeném uživateli, jeho úrovni přístupu a podobně. Pokud se klient může připojit k libovolnému serveru z clusteru, je třeba mezi jednotlivými servery informace o session [16] buď sdílet, nebo zajistit, že se stejný klient vždy připojí ke stejnému uzlu. V opačném případě by se klient mohl spojit se serverem, který o zahájené session neví, a tak přijít o přihlášení nebo stav nákupního košíku.



Obrázek 44: Přenos Session ID

Z hlediska skriptovacích jazyků pro programování internetových aplikací, session představuje množinu proměnných, které dovolují uchovávat hodnoty, které jim byly nastaveny po dobu připojení. V tomto případě se při znovunačtení stránky hodnoty neztratí.

5.3 Princip řešení

Po bližší analýze problémů, které souvisí s touto problematikou, patří k nejpřijatelnějšímu řešení realizace pomocí internetového prohlížeče.

Využití nástrojů prohlížeče, který má již v sobě implementovány tyto funkce, zajistí správu přihlašování přes protokol HTTPS, a také pozdější udržení SID, která je ze strany klienta zasílána v podobě cookie.

Samotný proces provádění automatických příhozů nebo přihlášení na internetové stránky je třeba provést pomocí skriptu, který je spouštěn v rámci prohlížeče. Tento skript je zakomponován v přídavném pluginu a je prováděn při načítání stránky. Schopnosti tohoto řešení jsou parsování stránky a zjištění potřebných informací jako čas ukončení aukce, současná výše částky v dražbě nebo odeslání formuláře s přesně definovanými hodnotami.

Některé internetové aukce také nabízejí možnost automatických příhozů jako nabídku jejich dalších služeb. Aktivace této služby bývá zpoplatněna. V poměru platby za poskytnutí služby automatických příhozů a výše návratnosti v prováděných dražbách není tato volba příliš úměrná.

Obě řešení mají za cíl usnadnit práci uživateli a minimalizovat čas strávený potřebnou kontrolou dražených položek. Tyto řešení zvyšují šanci na výhru v probíhané dražbě a snaží se eliminovat případy, že uživatel bude v posledních vteřinách přeskočen vyšším příhozem a o položku přijde.

6. Závěr

Výsledkem této bakalářské práce je plně funkční program, který je navržen pro zpracování internetových stránek a uložení požadovaných výsledků do cílového strukturovaného souboru typu XML. Zrealizována byla metoda postupného procházení internetových stránek, a také automatické generování nových URL.

Program je tedy schopen využít vlastností webových robotů a principu následování hypertextových odkazů. Byl aplikován algoritmus procházení směrem do hloubky s možností omezení maximálního relativního zanoření od vstupního bodu programu. Pro detailnější určení vyhledávaných odkazů byla vytvořena funkce k vytvoření podmínek, které musí nově nalezený hypertextový odkaz splňovat.

V metodě generování URL byly zvoleny číselné hodnoty jako parametry dosazované do vstupu. Tento způsob hledání internetových stránek byl rozšířen o možné kombinace těchto parametrů. Program podporuje 3 odlišné způsoby generování nových URL a tím byla rozšířena možnost jeho nasazení. Všechny metody byly optimalizovány pro efektivnější využití času, aby se zamezilo generování URL, které s velkou pravděpodobností nebudou vracet potřebné výsledky tak, aby nedošlo k předčasnému ukončení, pokud budou stále nacházeny stránky vyhovující podmínkám.

Hlavním přínosem a výhodou programu je použití na internetových stránkách s různou obecnou strukturou. K dosažení této funkce byl implementován model stromu, který slouží jako předloha, podle které je řízen způsob zpracování stránky, a jeho propojení s tabulkou atributů. Práce se stromem je velice intuitivní a snadná. Hlavním úkolem stromu je postupné vybírání požadovaných úseků textu na internetové stránce a předávání výsledků jednotlivým potomkům uzlů stromu. Vzájemné provázání s tabulkou atributů, jejíž záznamy jsou spojeny s konkrétními uzly stromu, zajišťuje správný zápis získaných dat do cílových souborů.

V druhé části bakalářské práce byla analyzována problematika simulování chování reálného uživatele v prostředí internetových aukcí. Důraz byl kladen na automatické přihlašování a provádění příhozů s předem definovanými podmínkami jako jsou výše příhozu nebo maximální akceptovatelná částka.

Výsledkem zjištění byla vybrána možnost použití internetových prohlížečů a jejich vestavěných funkcí. Tento způsob řeší problémy spojené s přihlášením přes zabezpečený protokol HTTPS, a také pozdější udržování relace mezi uživatelem a webovým serverem.

Samotné provedení příhozu je poté prováděno skriptem, který je na straně uživatele schopen odeslat formulář s přesně definovanými hodnotami nebo jím lze zjistit termín ukončení dražby.

Seznam použité literatury

1. *Web crawler - Wikipedia, the free encyclopedia* [online]. 2001- , 10 April 2009 [cit. 2009-04-11]. Funkce webového robota.
   Dostupný z WWW: <http://en.wikipedia.org/wiki/Web_crawler>.

2. LEVENE, Mark, POULOVASSILIS, Alexandra. *Web Dynamics : adapting to change in content, size, topology and use.* 2004th compl. edition. [s.l.] : Springer, c2004. Dostupný z WWW: <http://books.google.cz/books?id=q0qb5Vi02YAC&printsec=frontcover>. ISBN 978-3-540-406. Crawling algorithms, s. 163-167.

3. *Focused algorithm* [online]. 2007- , 2. April 2009 [cit. 2009-04-16]. Popis algoritmu Focused.
   Dostupný z WWW: <http://en.wikipedia.org/wiki/Focused_crawler>.

4. *RSS - Wikipedia, the free encyclo*pedia [online]. 2006- , 15.3.2009 [cit. 2009-04-06]. Princip fungování RSS.
   Dostupný z WWW: <http://cs.wikipedia.org/wiki/RSS>.

5. *Infragistics* [online]. 2009 [cit. 2009-04-28]. Grafická komponenta.
   Dostupný z WWW: <http://infragistics.com/>.

6. *Procházení grafu do hloubky - Wikipedia, the free encyclopedia* [online]. 2007- , 24.2.2009 [cit. 2009-04-08]. Popis prohledávání grafu do hloubky.
   Dostupný z WWW: <http://cs.wikipedia.org/wiki/Prohledávání_do_hloubky>.

7. *UTF-8 - Wikipedia, the free encyclopedia* [online]. 2004- , 30.1.2009 [cit. 2009-04-08]. Popis kódování UTF-8. Dostupný z WWW: <http://cs.wikipedia.org/w/index.php?title=UTF-8&action=history>.

8. *XSLT - Wikipedia, the free encyclopedia* [online]. 2006- , 5.2.2009 [cit. 2009-04-08]. Rozbor XSL transformace.
   Dostupný z WWW: <http://cs.wikipedia.org/wiki/XSLT>.

9. *Aurea Numismatika* [online]. 2002 [cit. 2009-04-16]. Internetová aukce.
   Dostupný z WWW: <http://www.aurea.cz/>.

10. *Coinarchives* [online]. 2009 [cit. 2009-04-16]. Internetová aukce.
    Dostupný z WWW: <http://www.coinarchives.com/>.

11. *Sixbid* [online]. 2009 [cit. 2009-04-16]. Internetová aukce.
Dostupný z WWW: <http://www.sixbid.com/>.

12. *Stack's* [online]. 2001-2009 [cit. 2009-04-16]. Internetová aukce.
Dostupný z WWW: <http://www.stacks.com/>.

13. *eBay* [online]. 1995-2009 [cit. 2009-04-16]. Aukční portál.
Dostupný z WWW: <http://www.ebay.com/>.

14. *Aukro* [online]. 2009 [cit. 2009-04-16]. Aukční portál.
Dostupný z WWW: <http://www.aukro.cz/>.

15. *HTTPS - Wikipedia, the free encyclopedia* [online]. 2006- , 3.4.2009 [cit. 2009-04-08]. Zabezpečený protokol HTTPS.
Dostupný z WWW: <http://cs.wikipedia.org/wiki/HTTPS>.

16. *Session - Wikipedia, the free encyclopedia* [online]. 2007- , 18.1.2009 [cit. 2009-04-08]. Popis fungování Session.
Dostupný z WWW: <http://cs.wikipedia.org/wiki/Session>.

Přílohy

A. Obsah přiloženého CD

Adresář	Obsah
/text/	Elektronická verze tohoto dokumentu
/aukce/aurea/	Vzorová data z aukce Aurea Numismatika
/aukce/coinarchives/	Vzorová data z aukce CoinArchives
/program/	Zdrojové kódy vytvořeného programu
/instalace/	Instalační soubory programu
/help/	Uživatelská nápověda
/šablona/aurea/	Nastavení programu pro aukci Aurea Numismatika
/šablona/sixbid/	Nastavení programu pro aukci Sixbid
/šablona/stacks/	Nastavení programu pro aukci Stack's
/šablona/coinarchives/	Nastavení programu pro aukci Coinarchives
/infragistics/	Instalační soubory použité grafické komponenty

B. Uživatelský manuál

Stahování online dat

Program se používá k získání požadovaných dat a uložení výsledku v jednotné formě. Program slouží jako datová pumpa, kde výsledný soubor je typu XML s jednotnou strukturou.

Výhody programu jsou:

1. Jednoduchost a přehlednost
2. Dynamická změna nastavení
3. Přívětivé uživatelské prostředí

Jak nainstalovat program a začít jej používat je vysvětleno v sekci Začínáme.
Nastavení jednotlivých částí programu je vysvětleno v sekci Práce s programem.

Začínáme

Podívejte se na tyto stránky pro informace o Instalaci a Spuštění programu.

- Instalace
- Spuštění programu

Instalace

Systémové požadavky

Windows Installer 3.1
.NET Framework 3.5 SP1
Windows XP/Vista

Instalace

Pro nainstalování programu spusťte soubor "setup.exe". V případě chyby při instalaci potřebných doplňků budete muset doplněk nainstalovat ručně.
Instalátor pro .NET Framework 3.5 SP1 naleznete zde.
Instalátor pro Windows Installer 3.1 naleznete zde.

Pro odinstalování programu použijte následující postup:

1. Klikněte na tlačítko [Start].
2. Vyberte Nastavení -> Ovládací panely.
3. Dvakrát klikněte na tlačítko Přidat nebo odebrat programy.
4. Pak vyberte program "Stahování online dat" ze seznamu.
5. Klikněte na tlačítko [Odinstalovat nebo ...] pro odebrání programu ze systému.

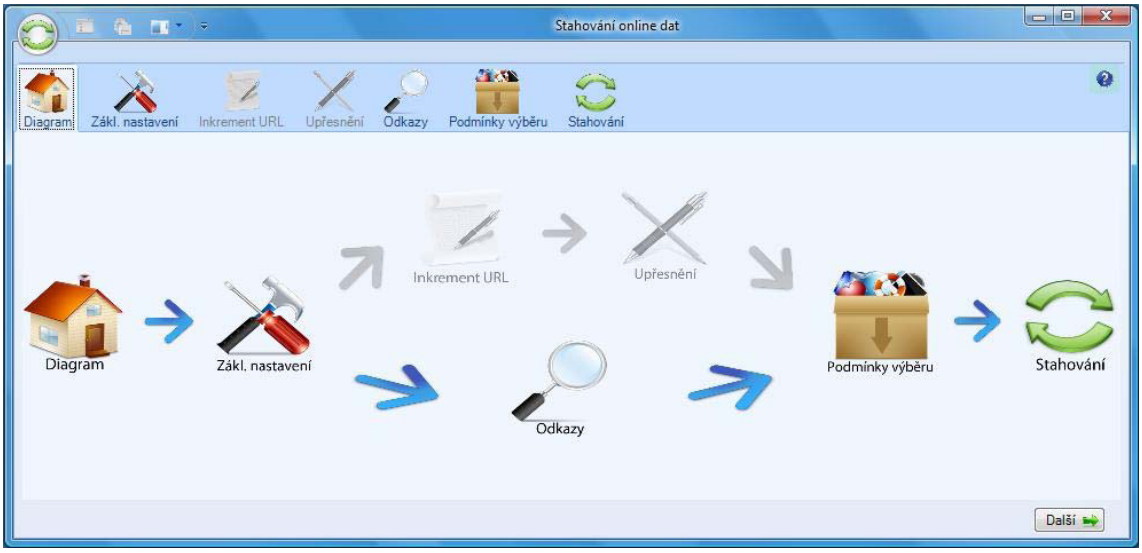
Spuštění programu

Chcete-li začít pracovat s programem, klikněte na ikonu Stahování online dat, kterou najdete v nabídce Start -> Programy -> VŠB - Technická univerzita Ostrava -> "Stahování online dat" nebo klikněte na ikonu "Stahování online dat", která se vám vytvořila na ploše.

Diagram

Graficky zobrazuje průběh postupu nastavení programu. Při kliknutí na ikonu zastupující záložku se přesunete do požadované části programu. Zašednuté a neaktivní ikony naznačují, že daná oblast programu ještě nebyla nastavena nebo k ní nemáte přístup. Po nastavení, kdy je program schopen provést požadované operace se zpřístupní také ikona pro "Stahování" a zvýrazní se průběh jednotlivých nastavení programu.

Obrázek níže zobrazuje způsob nastavení, kdy program prochází jednotlivé hypertextové odkazy.



Základní nastavení

Výběr a uložení aukce:

Zde se nastavují základní informace nezbytné pro běh programu.

V horní části si po stisknutí tlačítka "Procházet" vyberete složku na disku, do které chcete požadované soubory uložit.

Pod touto položkou se nachází místo, kde zadáte počáteční URL, ze které bude program hledat další hypertextové odkazy nebo generovat nová URL a zaškrtnete možnost "Inkrementovat URL aukce" pokud chcete, aby program sám generoval další URL.

Při zatržení možnosti "Inkrementovat URL" bude vaše další možnost pro nastavení programu záložka Inkrement URL. V opačném případě to bude záložka Odkazy.

Zpracování obrázků:

Při zvolení možnosti "Stahovat obrázky" se budou na disk ukládat i obrázky podle vámi zadaných kritérií a zpřístupní se vám možnost "Ořezávat linky obrázků".

Označením "Ořezávat linky obrázků" docílíte toho, že veškeré stažené soubory se budou nacházet ve stejném adresáři jako cílový soubor typu XML

Zatrhnutím "Ukládat linky obrázků" budete požadovat, aby se původní URL obrázků v absolutním tvaru uložily do výsledného souboru.

Kontrola kódování:

V této oblasti si vyberete, jestli chcete zjišťovat kódování, ve kterém je stránka napsána, pouze na první vámi zadané URL a toto kódování aplikovat i na zbylé stránky nebo jestli chcete kódování stránky kontrolovat vždy.

Archivování souborů:

Při zvolení možnosti "Přidat do archívu 7-Zip" se veškerá získaná data zabalí do archívu, který se uloží do vámi zvolené cílové složky.

Pokud vyberete i možnost "Nechat pouze archivované soubory", tak celkový stažený obsah bude smazán a v cílové složce zůstane pouze požadovaný archív.

Zde je náhled, jak může být tato sekce vyplněna:

Inkrement URL

V této části nastavujete, jakým způsobem se budou generovat nová URL.

Označení dat:

Zde se vám překopíruje URL ze záložky "Základní nastavení" a v něm označíte část URL příslušnými značkami, které odpovídají úseku "Lot", "ID Aukce"nebo "Lot ID".

Vámi označená část bude v průběhu generování URL nahrazována čísly z rozsahu, které nastavíte pro příslušné značky.

Podmínka parsování stránky:

Zde zapište text, který se musí nebo naopak nesmí na stránce objevit, aby byla dále zpracovávána a program se pokusil získat potřebná data.

Výběr algoritmu

Program nabízí 3 možnosti pro generování URL:

1. Lot
2. ID Aukce + Lot
3. Lot ID + Lot

Lot:

Tato možnost pouze doplňuje jednotlivá čísla ze zadaného rozsahu za text označený mezi značkami pro "Lot".

Pokud je "Zarovnat na počet znaků" nastaven na hodnotu 0, tak se neprovádí žádná akce. V opačném případě se chybějící část znaků doplňuje symbolem "nula".

Např. při nastavení zarovnání na 5 znaků se z čísla 38 -> 00038.

V části "Od:" "Do:" se zadává nejmenší respektive největší číslo pro dosazení do URL.

"Tolerance špatných stránek" označuje, kolik generovaných URL může nevyhovovat podmínkám pro parsování stránky a v případě dosažení daného počtu se program ukončí nebo pokud je tak nastaveno skočí o zadaný počet čísel.

"Při správném projdění ..." nastavujete o kolik čísel se má zadaný rozsah zvětšit, pokud byl celý rozsah zkontrolován a program stále nachází stránky, které splňují zadané podmínky pro parsování.

"Při nesprávném projdění ..." vyplňujete, o kolik čísel vpřed se má počítadlo posunout, pokud program dosáhl maximálního počtu stránek, které za sebou nesplňovaly podmínky pro parsování.

ID Aukce (postupný inkrement):

V tomto případě se označená část URL nahrazuje způsobem, kdy se počáteční číslo rozsahu pro "ID Aukce" dosadí do URL a k němu se dosazuje celý rozsah nastavení "Lotu".

Po skončení rozsahu "Lotu" se číslo "ID Aukce" o 1 zvýší a opět se dosazuje celý rozsah nastavení "Lotu". Program pokračuje dokud se "ID Aukce" nedostane na konec rozsahu.

Pokud je "Zarovnat na počet znaků" nastaven na hodnotu 0, tak se neprovádí žádná akce. V opačném případě se chybějící část znaků doplňuje symbolem "nula".

Např. při nastavení zarovnání na 5 znaků se z čísla 38 -> 00038.

V části "Od:" "Do:" se zadává nejmenší respektive největší číslo pro dosazení do URL.

"Tolerance prázdných aukcí" označuje, kolik po sobě jdoucích čísel v "ID Aukce" nesmí obsahovat pouze stránky, které nevyhovují podmínce pro parsování. V případě dosažení zadaného počtu se generování URL ukončí.

"Při správném projdění ..." určujete o kolik se má rozsah pro "ID Aukce" rozšířit, pokud poslední generovaná URL splňovala podmínky pro parsování.

Lot ID (současný inkrement):

Při této možnosti se dosazovaná čísla z rozsahu "Lot ID" i "Lot" zvyšují současně.

Pokud je "Zarovnat na počet znaků" nastaven na hodnotu 0, tak se neprovádí žádná akce. V opačném případě se chybějící část znaků doplňuje symbolem "nula".

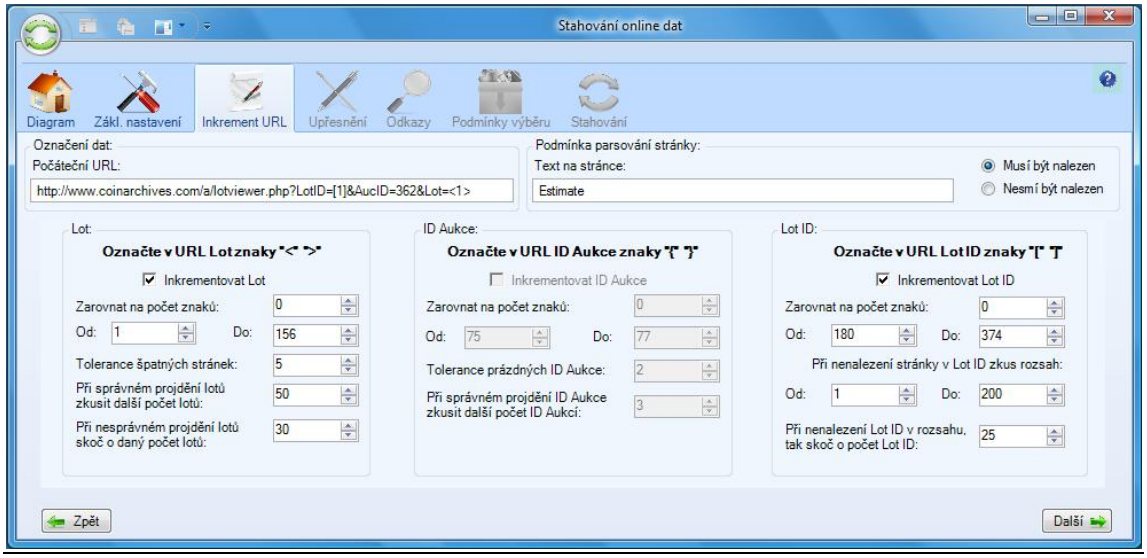
Např. při nastavení zarovnání na 5 znaků se z čísla 38 -> 00038.

V části "Od:" "Do:" se zadává nejmenší respektive největší číslo pro dosazení do URL.

Při nenalezení stránky v "Lot ID" zkus rozsah: Je to rozsah čísel, která se budou dosazovat místo "Lotu", pokud byla dosažena "Tolerance špatných stránek" z "Lotu" a číslo "Lot ID" zůstane po celou dobu nezměněno. Po nalezení stránky, která splňuje podmínky pro parsování se opět začnou jak "Lot ID" tak "Lot" zvyšovat současně. Při nenalezení stránky v daném rozsahu se číslo "Lot ID" zvýší o počet zadaný v oblasti pro skok Lot ID.

"Při nenalezení Lot ID ..." o kolik čísel se má Lot ID zvýšit a dále pokračovat v současném zvyšování s "Lotem".

Obrázek níže ukazuje nastavení pro způsob (současný inkrement).



Upřesnění

V této sekci zadáváte doplňující informace pro generování URL.

Pokud jste na záložce "Zákl. nastavení" zatrhli volbu "Stahovat obrázky", tak se vám zpřístupní možnost pro vyplnění podmínek pro stahování obrázků.

Podmínky stahování obrázků:

Zapsáním textu, který musí URL obrázku obsahovat určujete podmínky, které musí obrázek splňovat, aby byl stažen. V seznamu podmínek pro stahování obrázků může být více záznamů pro různé druhy podmínek.

Seznam podmínek lze měnit podle potřeby.

Zpracování dat ze souboru:

Při zvolení volby "Načíst data ze souboru" se zpřístupní možnost pro vybrání souboru na disku, který obsahuje posloupnost čísel, která se bude dosazovat místo označené části v URL ze záložky "Inkrement URL".

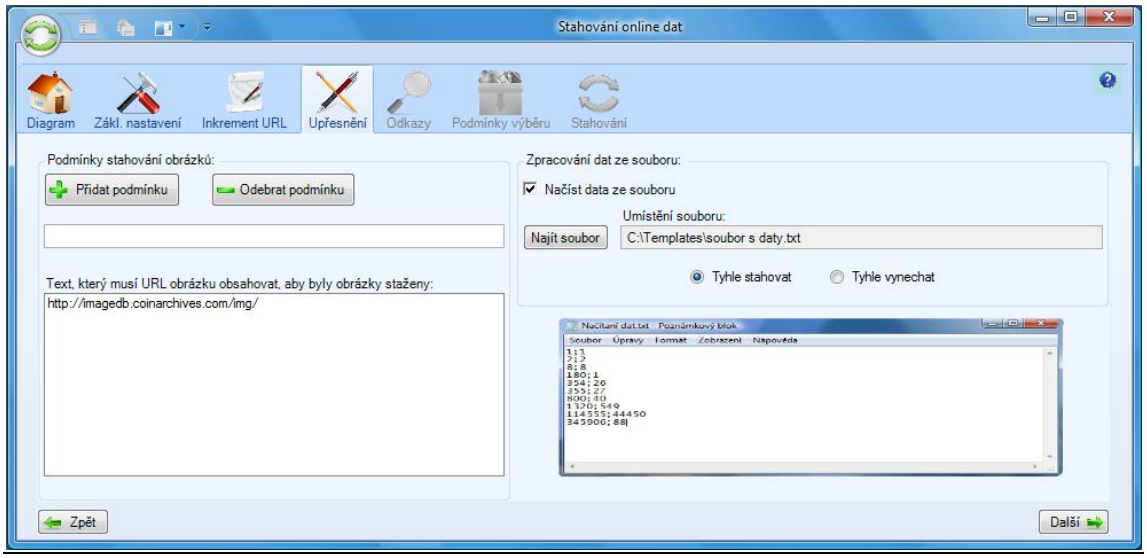
Dosazování čísel bude také probíhat podle zvoleného způsobu ze záložky "Inkrement URL".

Při možnosti generování URL pomocí "Lot" musí soubor obsahovat jedno číslo na řádku a další číslo na novém řádku.

Při možnosti generování URL pomocí "ID Aukce + Lot" nebo "Lot ID + Lot" musí soubor obsahovat dva čísla na jednom řádku oddělená středníkem (pozn. značka středníku ";"). Další dvojice čísel se nachází opět na novém řádku.

Čísla jsou zadána ve výše uvedeném pořadí.

Obrázek názorně ukazuje nastavení záložky.



Odkazy

Na této záložce zadáváte podmínky, podle kterých se budou stránky procházet pomocí hypertextových odkazů.

Maximální hloubka zanoření:

Určuje, jaké může být největší zanoření v rámci procházení hypertextových odkazů. Při dosažení maximální hloubky se již nehledají stránky s hlubším zanořením, ale zpětně se kontrolují nalezené stránky a opět se prohledávají do určené hloubky zanoření do doby, než jsou zkontrolovány všechny nalezené odkazy splňující podmínky.

Podmínky stahování obrázků a parsování stránek:

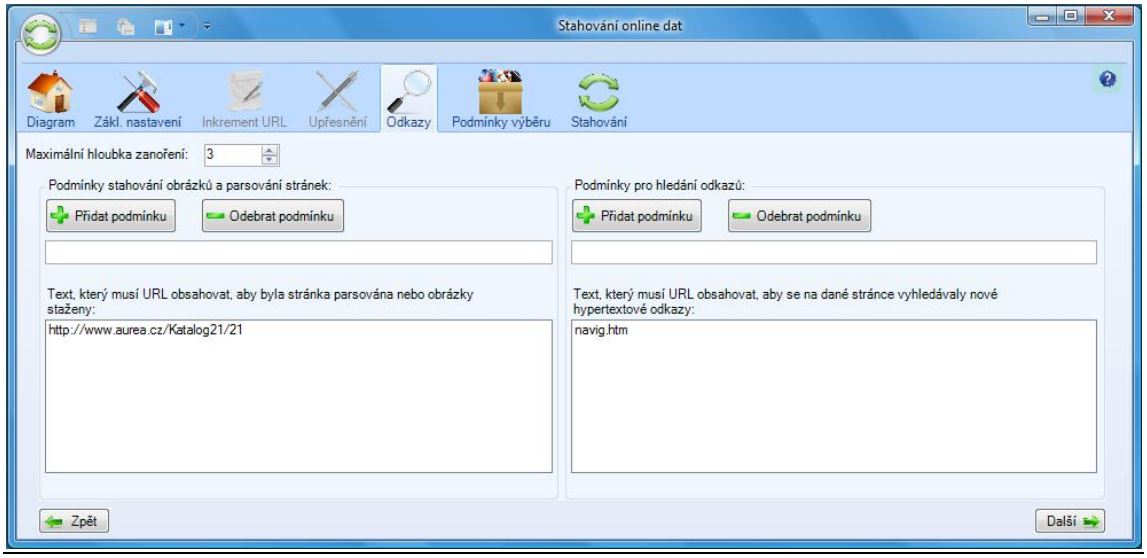
Zadáváte text, který musí URL obrázku obsahovat, aby byl stažen případně, aby stránka byla parsována.

Můžete zadat více podmínek, které se mohou lišit v případě podmínek pro stránky nebo obrázky. Tyto podmínky jsou řazeny do seznamu.

Podmínky pro hledání odkazů:

Zadáváte text, který musí URL stránky obsahovat, aby na ní byly vyhledány další hypertextové odkazy, které budou splňovat podmínku pro parsování stránky nebo podmínku hledání dalších odkazů.

Obrázek zobrazuje příklad nastavení záložky.



Podmínky výběru

Zde zadáváte způsob, jakým bude stránka, která se má parsovat zpracována. Slouží k tomu především strom, podle kterého se zjišťuje, jakým způsobem se bude obsah stránky členit a tabulka atributů, ze které se vybírají názvy jednotlivých uzlů stromu.

Strom pro parsování:

Strom je realizován způsobem, kdy uzly na stejné úrovni postupně odebírají obsah stránky podle zadaných počátečních a koncových značek. Zato uzlům, které jsou zanořeny hlouběji, předávají zbylý obsah stránky ke zpracování.

Uzly, které již nemají nikoho, komu by mohly předat zbylý text, zapisují výsledná data do koncového souboru mezi značky, které nesou jejich název.

Každý uzel má v sobě uložený jednoznačný počáteční a koncový tag, podle kterých vyhledá text v obsahu stránky a zpracuje jej.

Další způsob práce s jednotlivými uzly naleznete v sekci Strom.

Tabulka atributů:

Obsahuje názvy jednotlivých uzlů, které mohou být přidány do stromu.

Tabulka také určuje, jestli je daný uzel povinný. V případě, že ano a v průběhu parsování stránky došlo k tomu, že nebylo podle zadaných podmínek nic nalezeno, tak se do koncového souboru zapíše text, který je obsažen v tabulce ve sloupci "Náhrada zápisu". V druhém případě zůstane obsah zapsaných dat prázdný.

Upozornění:

Pro korektní vytvoření šablony k zobrazení souboru je doporučeno pro uzly, které budou ukládat obrázek zapsat text "".

Pozn. bez prvních a posledních uvozovek

Do sloupce "Náhrada zápisu" nevpisujte žádné HTML značky, které by mohl způsobit nekorektnost výsledného souboru.

Další způsob práce s tabulkou naleznete v sekci Atributy.

Setřídit výsledný soubor:

Ve výběru atributů můžete zvolit jeden z uzlů, které budou ukládat data do koncového souboru a podle něj výsledný soubor setřídit.

Ve výběru Typ zvolte, jestli bude konkrétní uzel obsahovat v koncovém souboru číslo nebo text.

Rozdělit výsledný soubor:

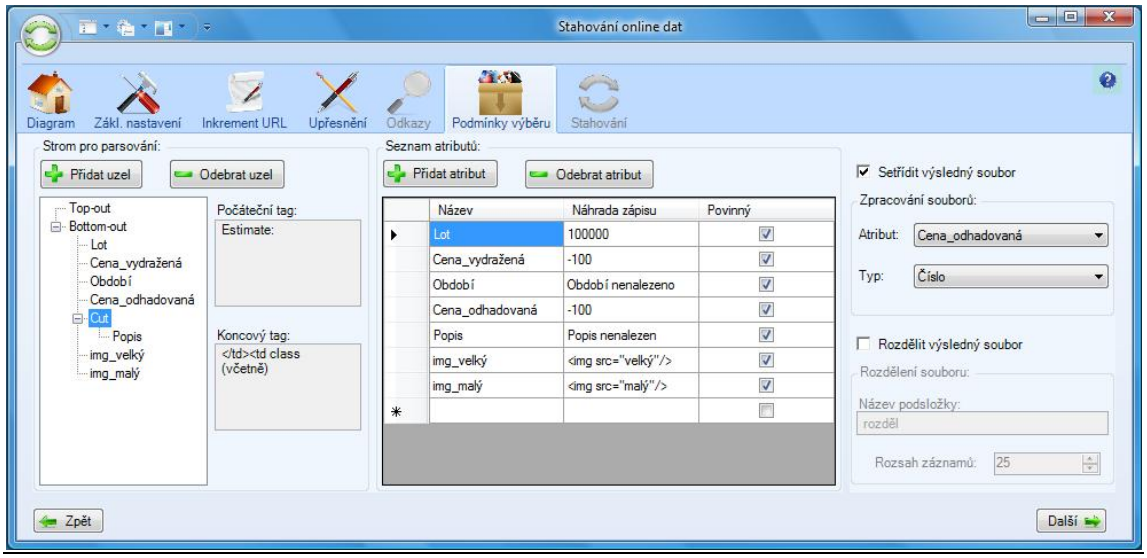
Rozsah záznamu určuje, kolik položek bude obsaženo v jednom souboru. V případě dosažení počtu položek v souboru se vytvoří nový soubor o stejném rozsahu položek a další data se budou ukládat do již nově vytvořeného souboru.

Pro každý nový koncový soubor se vytvoří nová složka pojmenovaná ze zadaného textu pro "Název podsložky" a za to se přiřadí rozsah položek, které složka obsahuje.

Výsledný soubor bude pojmenován stejným způsobem.

Např. je-li zadán název podsložky "Aurea" a rozsah je nastaven na 500, tak výsledný soubor se bude jmenovat "Aurea1-500.xml".

Můžete si prohlédnout obrázek případného nastavení této záložky.



Stahování

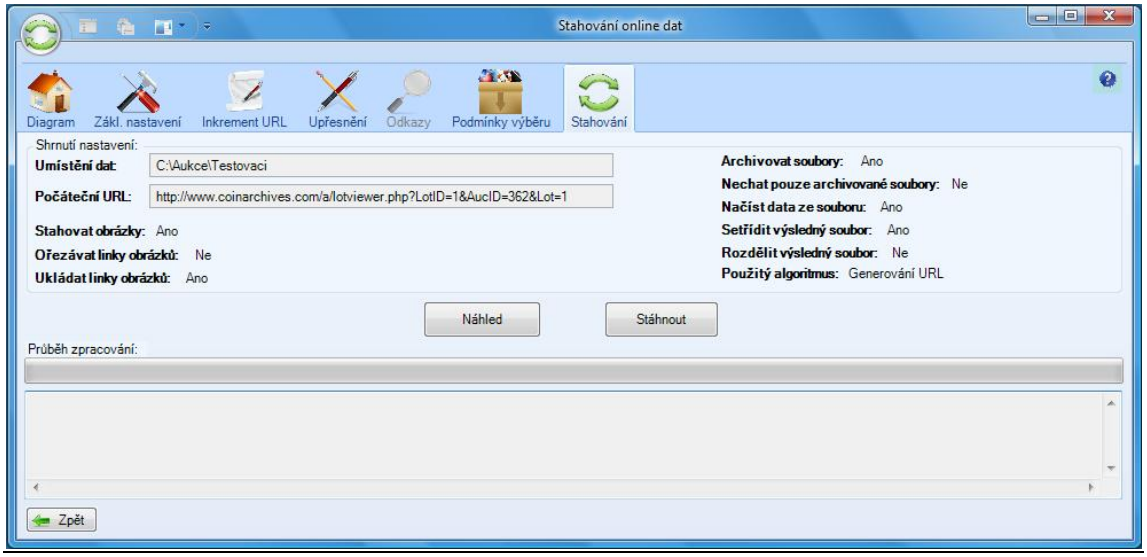
Tato záložka slouží k informativním účelům a ke spuštění samotného získávání potřebných dat.

Před započetím samotného stahování dat máte možnost si zobrazit náhled, jak bude výsledný soubor vypadat a jak bude formátován. Tato funkce vám pomůže si upřesnit nastavení programu a pomůže zkontrolovat nastavení parsování stránky.

Tlačítko "Náhled" tedy zobrazuje, jak bude výsledný soubor vypadat a pro zobrazení je využit program, který máte nastaven pro zobrazení souborů typu XML. Výchozí nastavení systému je IE.

Tlačítko "Stáhnout" spouští program se zadaným nastavením.

Obrázek ukazuje vzhled záložky "Stahování".



Menu

- Šablony
- Ostatní

Šablony

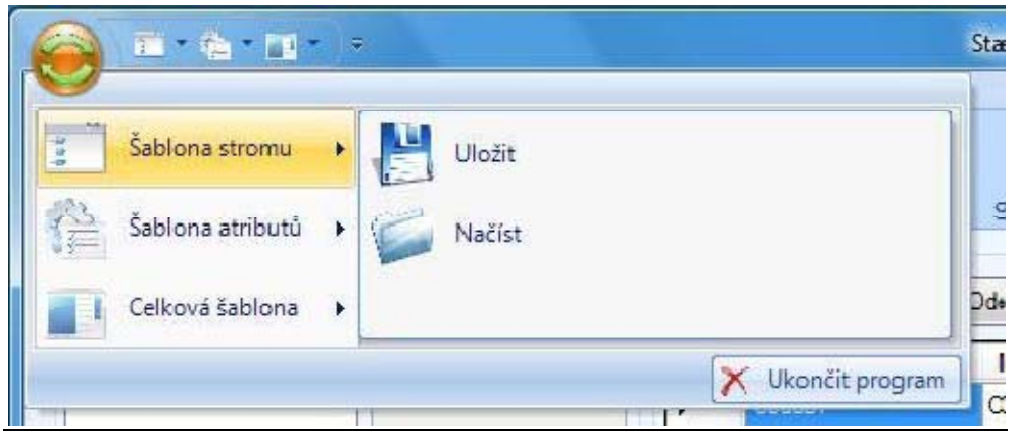
Šablona slouží k usnadnění práce s programem. Pro pozdější použití je možno si uložit šablonu Stromu, šablonu Tabulky atributů a šablonu celého nastavení programu.

Při spuštění programu stačí pouze vybrat požadovaný soubor s nastavením a použít jej.

Šablona stromu:

V šabloně je uloženo nastavení jednotlivých uzlů stromu se všemi dodatečnými informacemi. Ukládání a načítání šablon stromu naleznete v menu na těchto místech:

Práce se šablonou v hlavním menu.



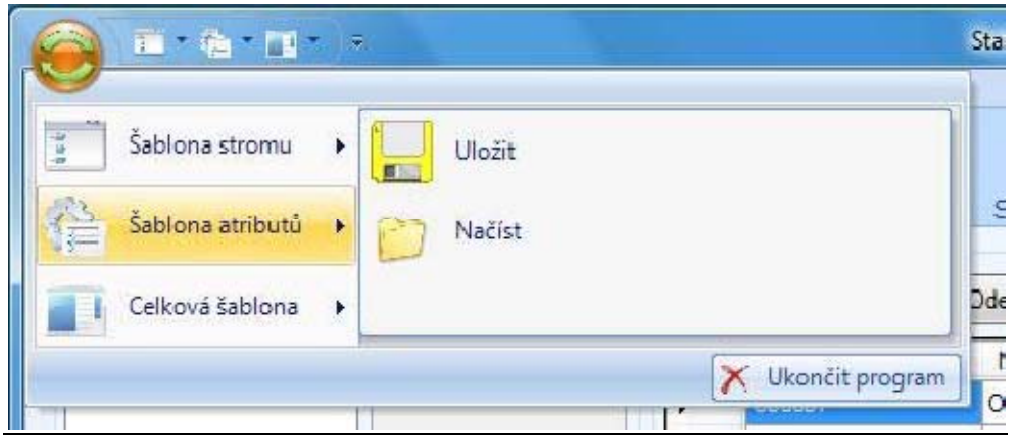
Práce se šablonou z horního panelu.

Možnost "Uložit" a "Načíst" se zpřístupní po přepnutí na záložku Podmínky výběru.

Šablona atributů:

V šabloně je uložena celá tabulka atributů se zadanými hodnotami.
Ukládání a načítání šablon tabulky atributů naleznete v menu na těchto místech:

Práce se šablonou v hlavním menu.



Práce se šablonou z horního panelu.

Možnost "Uložit" a "Načíst" se zpřístupní po přepnutí na záložku Podmínky výběru.

Celková šablona:

V šabloně jsou uloženy informace o všech nastaveních programu, které jste provedli.

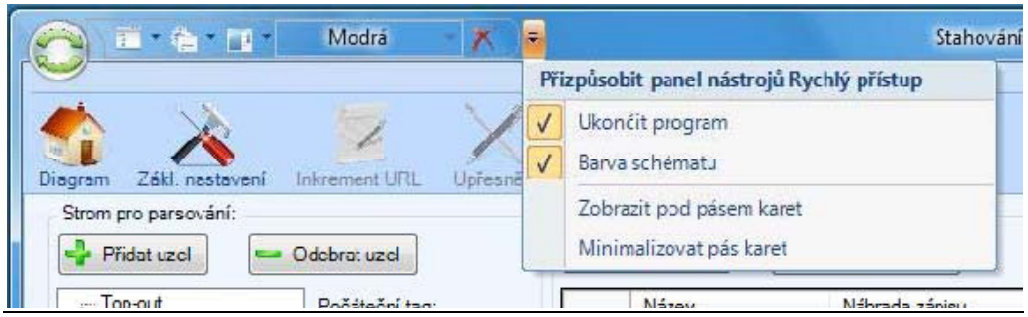
Ukládání a načítání šablony pro celkové nastavení programu naleznete v menu na těchto místech:

Práce se šablonou v hlavním menu.

Práce se šablonou z horního panelu.

Ostatní

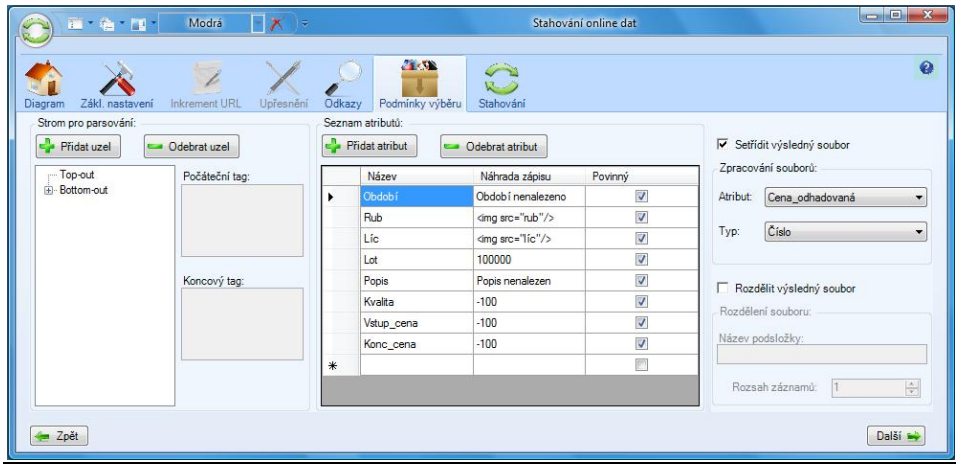
Menu nabízí možnost rozšířit nabídku panelu nástrojů Rychlý přístup o "Barvu schématu" a "Ukončení programu".



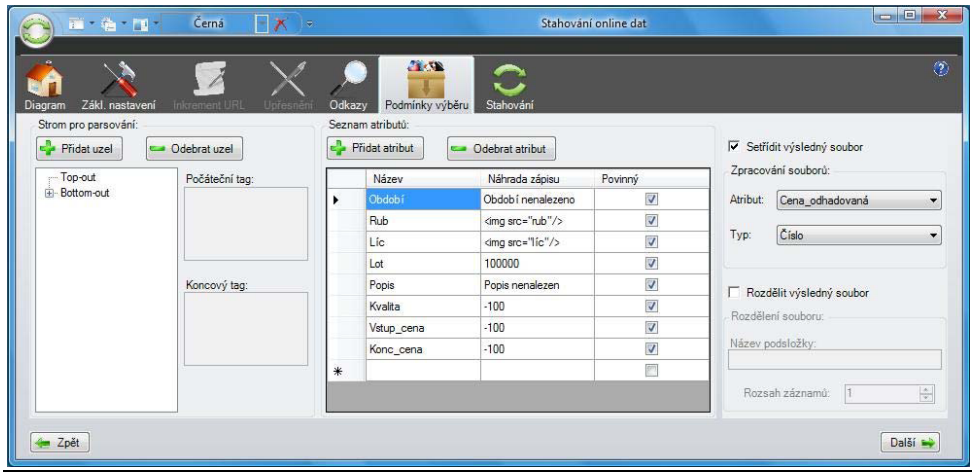
Barva schématu:

Po přidání "Barva schématu" na panel nástrojů si můžete vybrat ze tří barevných provedení vzhledu programu:

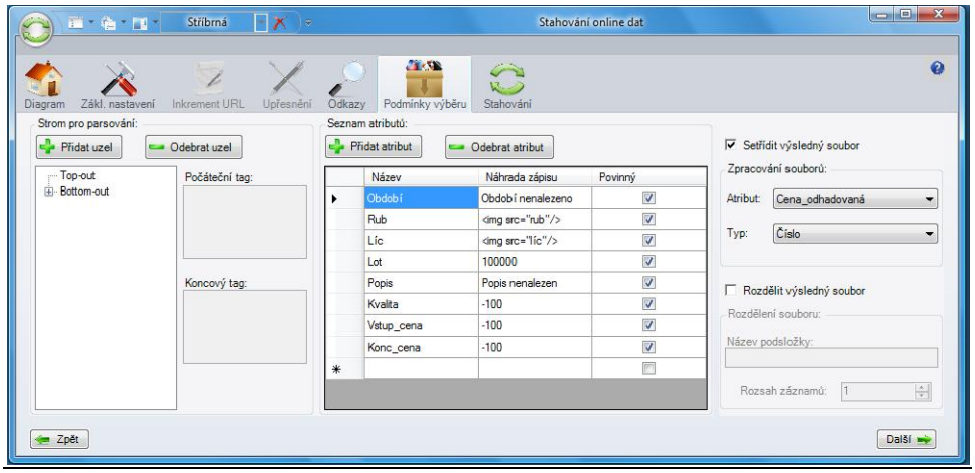
- **Modrá**



- **Černá**



- **Stříbrná**



Dialogy

Dialogy pro detailnější nastavení programu.

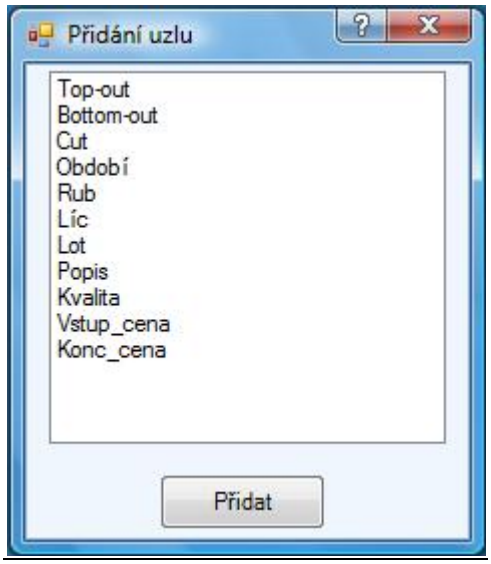
- Strom
- Atributy

Strom

Nastavení a využití stromu se nachází v sekci Podmínky výběru.

Přidání uzlu:

Přidání uzlu do stromu se provádí vybráním položky ze seznamu a stisknutím tlačítka "Přidat".



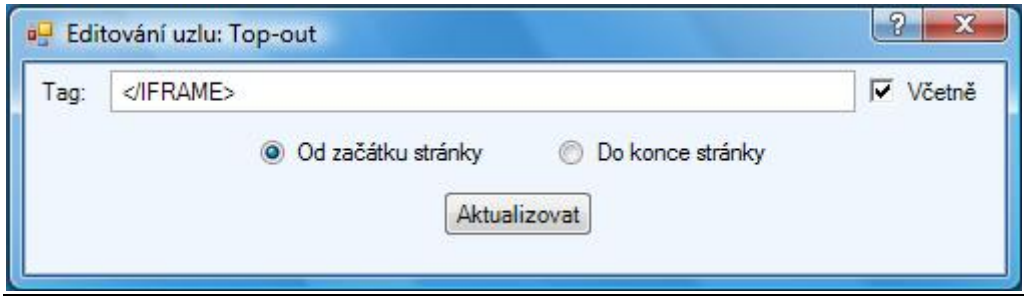
Editování uzlu:

Každý uzel stromu musí obsahovat počáteční a koncový tag a označení, zda se má obsah stránky zpracovávat včetně tagu nebo až za ním respektive před ním.

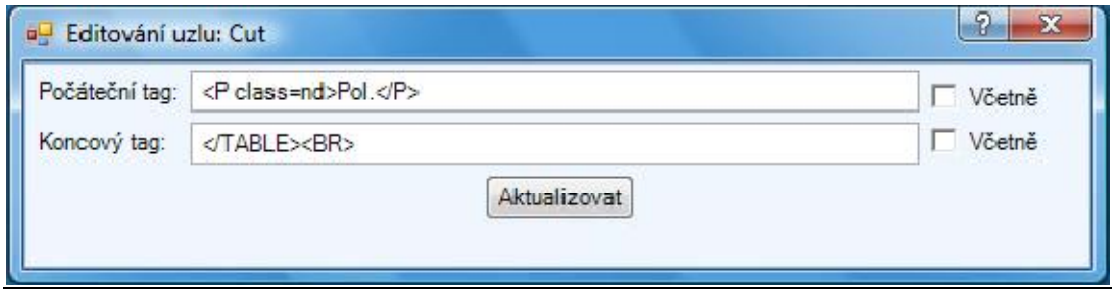
Uzly "Top-out" a "Bottom-out" obsahují nastavení, podle kterého lze odebrat text od počátků stránky do požadovaného tagu případně od zadaného tagu do konce stránky.

Přidávat a upravovat nastavení uzlu se vyvolá po stisknutí pravého tlačítka nad daným uzlem.

Okno pro editování uzlu "Top-out" a "Bottom-out"



Okno pro editování ostatních uzlů



Atributy

Nastavení a využití atributu se nachází v sekci Podmínky výběru.

Tabulku atributů lze modifikovat přímo z hlavního menu po stisknutí příslušné buňky a vepsání požadovaných dat nebo také stisknutím tlačítka "Přidat atribut" a tím vyvolat okno pro vložení nového řádku do tabulky.

Sloupec "Název" nesmí být prázdný a také nesmí obsahovat dvakrát stejný text.

Okno pro přidání nového řádku tabulky